SARS VIRUS NUCLEOTIDE AND AMINO ACID SEQUENCES AND USES THEREOF

Field of the Invention

The invention is in the field of virology. More specifically, the invention is in the field of coronaviruses.

Background of the Invention

Severe acute respiratory syndrome (SARS), a worldwide outbreak of atypical pneumonia with an overall mortality rate of about 3 to 6%, has been attributed to a coronavirus following tests of causation according to Koch's postulates, including monkey inoculation (R. Munch, *Microbes Infect* 5, 69-74, Jan. 2003). The coronaviruses are members of a family of enveloped viruses that replicate in the cytoplasm of animal host cells (B. N. Fields et al., *Fields virology*, Lippincott Williams & Wilkins, Philadelphia, 4th ed., 2001). They are distinguished by the presence of a single-stranded plus sense RNA genome, approximately 30 kb in length, that has a 5' cap structure and 3' polyA tract. Hence the genome is essentially a very large mRNA. Upon infection of an appropriate host cell, the 5'-most open reading frame (ORF) of the viral genome is translated into a large polyprotein that is cleaved by viral-encoded proteases to release several nonstructural proteins including an RNA-dependent RNA polymerase (Pol) and an ATPase helicase (Hel). These proteins in turn are responsible for replicating the viral genome as well as generating nested transcripts that are used in the synthesis of the viral proteins. The mechanism by which these subgenomic mRNAs are made is not fully understood, however transcription regulating sequences (TRSs) at the 5'end of each gene may represent signals that regulate the discontinuous transcription of subgenomic mRNAs (sgmRNAs). The TRSs include a partially conserved core sequence (CS) that in some coronaviruses is 5'-CUAAAC-3'. Two major models have been proposed to explain the discontinuous transcription in coronaviruses and arterioviruses (M.M.C.Lai, D. Cavanagh, Adv Virus Res. 48,1(1997); S. G. Sawicki, D.L. Sawicki,Adv.Exp. Med Biol.440,215(1998)). The

discovery of transcriptionally active, subgenomic-size minus strands containing the antileader sequence and transcription intermediates active in the synthesis of mRNAs (D. L. Sawicki et al., J. Gen Virol 82,386 (2001); S. G. Sawicki, D.L. Sawicki, J. Virol. 64,1050 (1990); M. Schaad, R.S.J. Baric,J. Virol.68,8169(1994); P. B. Sethna et al., Proc. Natl. Acad. Sci. U.S.A. 86,5626 (1989)) favors the model of discontinuous transcription during the minus strand synthesis(S. G. Sawicki, D.L. Sawicki,Adv.Exp. Med Biol.440,215(1998)).

The coronaviral membrane proteins, including the major proteins S (Spike) and M (Membrane), are inserted into the endoplasmic reticulum Golgi intermediate compartment (ERGIC) while full length replicated RNA (+ strands) assemble with the N (nucleocapsid) protein. This RNA-protein complex then associates with the M protein embedded in the membranes of the ER and virus particles form as the nucleocapsid complex buds into the ER. The virus then migrates through the Golgi complex and eventually exits the cell, likely by exocytosis (B. N. Fields et al., *Fields virology*, Lippincott Williams & Wilkins, Philadelphia, 4th ed., 2001). The site of viral attachment to the host cell resides within the S protein.

The coronaviruses include a large number of viruses that infect different animal species. The predominant diseases associated with these viruses are respiratory and enteric infections, although hepatic and neurological diseases also occur with some viruses. Coronaviruses are divided into three serotypes, Types I, II and III. Phylogenetic analysis of coronavirus sequences also identifies three main classes of these viruses, corresponding to each of the three serotypes. Type II coronaviruses contain a hemagglutinin esterase (HE) gene homologous to that of Influenza C virus. It is presumed that the precursor of the Type II coronaviruses acquired HE as a result of a recombination event within a doubly infected host cell.

In view of the rapid worldwide dissemination of SARS, which has the potential of creating a pandemic, along with its alarming morbidity and mortality rates, it would be useful to have a better understanding of this coronavirus agent at the molecular level to provide diagnostics, vaccines, and therapeutics, and to support public health control measures.

Summary of the Invention

In general, the invention provides the genomic sequence of a novel coronavirus, the SARS virus, and provides novel nucleic acid molecules encoding novel proteins that may be used, for example, for the diagnosis or therapy of a variety of SARS virus-related disorders.

In one aspect, the invention provides a substantially pure SARS virus nucleic acid molecule or fragment thereof, for example, a genomic RNA or DNA, cDNA, synthetic DNA, or mRNA molecule. In some embodiments, the nucleic acid molecule includes a sequence substantially identical to any of the sequences of SEQ ID NOs: 1-13, 15-18, 20-30, 90-159, 208, 209. In some embodiments, the nucleic acid molecule includes a sequence from SEQ ID NO: 1, SEQ ID NO:2, or SEQ ID NO: 15 or a fragment of these sequences. In alternative embodiments, the nucleic acid molecule may include a sequence substantially identical to SEQ ID NO: 1, SEQ ID NO:2, or SEQ ID NO: 15, or a fragment thereof. In alternative embodiments, the nucleic acid molecule may include a s2m motif (for example, a s2m sequence substantially identical to any of the sequence of SEQ ID NOs: 16, 17, and 18), a leader sequence (for example, a sequence substantially identical to the sequence of SEQ ID NO: 3), or a transcriptional regulatory sequence (for example, a sequence substantially identical to any of the sequence of SEQ ID NOs: 4-13 and 20-30). In alternative embodiments, the nucleic acid molecule includes a sequence substantially identical to any of the sequences of nucleotides 265-13,398; 13,398-21,485; 21,492 – 25,259; 25,268 – 26,092; 25,689 – 26,153; 26,117 – 26,347; 26,398 – 27,063; 27,074 – 27,265; 27,273 – 27,641; 27,638 – 27,772; 27,779 – 27,898; 27,864 – 28,118; 28,120 – 29,388; 28,130 – 28,426; 28,583 – 28,795; and 29,590 – 29,621 of SEQ ID NO: 15. In alternative embodiments, the nucleic acid molecule may encode a polyprotein or a polypeptide. In alternative embodiments, the invention provides a nucleic acid molecule including a sequence complementary to a SARS virus nucleotide sequence.

In an alternative aspect, the invention provides a substantially pure SARS virus polypeptide or fragment thereof, for example, a polyprotein, glycoprotein (for example, a matrix glycoprotein that may include a sequence substantially identical to the sequence of SEQ ID NO: 34), a transmembrane protein (for example, a multitransmembrane protein, a type I transmembrane protein, or a type II

transmembrane protein), a RNA binding protein, or a viral envelope protein. In alternative embodiments, the invention provides a replicase 1a protein, replicase 1b protein, a spike glycoprotein, a small envelope protein, a matrix glycoprotein, or a nucleocapsid protein. In alternative embodiments, the invention provides a nucleic acid molecule encoding a SARS virus polypeptide. In alternative embodiments, the SARS virus polypeptide includes an identifiable signal sequence (for example, a signal sequence substantially identical to the sequence of SEQ ID NOs: 76 or 85), a transmembrane domain (for example, a transmembrane domain substantially identical to any of the sequences of SEQ ID NOs: 77-86), a transmembrane anchor, a transmembrane helix, an ATP-binding domain, a nuclear localization signal, a hydrophilic domain, (for example, a hydrophilic domain substantially identical to the sequence of SEQ ID NOs: 87), or a lysine-rich sequence (for example, a sequence substantially identical to the sequence of SEQ ID NO: 14). In alternative embodiments, the SARS virus polypeptide may include a sequence substantially identical to any of the sequences of SEQ ID NOs: 14, 33-36, 64-74, and 76-87.

In alternative embodiments, the invention provides a vector (for example, a gene therapy vector or a cloning vector) including a SARS virus nucleic acid molecule (for example, a molecule including a sequence substantially identical to any of the sequences of SEQ ID NOs: 1-13, 15-18, 20-30, 90-159, 208, 209), or a host cell (for example, a mammalian cell, a yeast, a bacterium, or a nematode cell) including the vector.

In alternative embodiments, the invention provides a nucleic acid molecule having substantial nucleotide sequence identity (for example, 30%, 40%, 50%, 60%, 70%, 80%, 90% or 100% complementarity) to a sequence encoding a SARS virus polypeptide or fragment thereof, for example where the fragment includes at least six amino acids, and where the nucleic acid molecule hybridizes under high stringency conditions to at least a portion of a SARS virus nucleic acid molecule.

In alternative embodiments, the invention provides a nucleic acid molecule having substantial nucleotide sequence identity (for example, 30%, 40%, 50%, 60%, 70%, 80%, 90% or 100% complementarity) to a SARS virus nucleotide sequence, for example where the nucleic acid molecule includes at least ten nucleotides, and where

the nucleic acid molecule hybridizes under high stringency conditions to at least a portion of a SARS virus nucleic acid molecule.

In alternative embodiments, the invention provides a nucleic acid molecule comprising a sequence that is antisense to a SARS virus nucleic acid molecule, or an antibody (for example, a neutralizing antibody) that specifically binds to a SARS virus polypeptide.

In alternative embodiments, the invention provides a method for detecting a SARS epitope, such as a virion or polypeptide in a sample, by contacting the sample with an antibody that specifically binds a SARS epitope, such as a virus polypeptide, and determining whether the antibody specifically binds to the polypeptide. In alternative embodiments, the invention provides a method for detecting a SARS virus genome, gene, or homolog or fragment thereof in a sample by contacting a SARS virus nucleic acid molecule, for example where the nucleic acid molecule includes at least ten nucleotides, with a preparation of genomic DNA from the sample, under hybridization conditions providing detection of DNA sequences having nucleotide sequence identity to a SARS virus nucleic acid molecule. In alternative embodiments, the invention provides a method of targeting a protein for secretion from a cell, by attaching a signal sequence from a SARS virus polypeptide to the protein, such that the protein is secreted from the cell.

In alternative aspects, the invention provides a method for eliciting an immune response in an animal, by identifying an animal infected with or at risk for infection with a SARS virus and administering a SARS virus polypeptide or fragment thereof or fragment thereof, or administering a SARS virus nucleic acid molecule encoding a SARS virus polypeptide or fragment thereof to the animal. In alternative embodiments, the administering results in the production of an antibody in the mammal, or results in the generation of cytotoxic or helper T-lymphocytes in the mammal.

In alternative embodiments, the invention provides a kit for detecting the presence of a SARS virus nucleic acid molecule or polypeptide in a sample, where the kit includes a SARS virus nucleic acid molecule, or an antibody that specifically binds a SARS virus polypeptide.

In alternative aspects the invention provides a method for treating or preventing a SARS virus infection by identifying an animal (e.g., a human) infected with or at risk

for infection with a SARS virus, and administering a SARS virus nucleic acid molecule or polypeptide, or administering a compound that inhibits pathogenicity or replication of a SARS virus, to the animal. In alternative embodiments, the invention provides the use of a SARS virus nucleic acid molecule or polypeptide for treating or preventing a SARS virus infection.

In alternative aspects the invention provides a method of identifying a compound for treating or preventing a SARS virus infection, by contacting sample including a SARS virus nucleic acid molecule or contacting a SARS virus polypeptide with the compound, where an increase or decrease in the expression or activity of the nucleic acid molecule or the polypeptide identifies a compound for treating or preventing a SARS virus infection.

In alternative aspects the invention provides a vaccine (e.g., a DNA vaccine) including a SARS virus nucleic acid molecule or polypeptide.

In alternative aspects the invention provides a microarray including a plurality of elements, wherein each element includes one or more distinct nucleic acid or amino acid sequences, and where the sequences are selected from a SARS virus nucleic acid molecule or polypeptide, or a antibody that specifically binds a SARS virus nucleic acid molecule or polypeptide.

In alternative aspects the invention provides a computer readable record (e.g., a database) including distinct SARS virus nucleic acid or amino acid sequences.

A "SARS virus" is a virus putatively belonging to the coronavirus family and identified as the causative agent for sudden acute respiratory syndrome (SARS). A SARS virus nucleic acid molecule may include a sequence substantially identical to the nucleotide sequences described herein or fragments thereof. A SARS virus polypeptide may include a sequence substantially identical to a sequence encoded by a SARS virus nucleic acid molecule, or may include a sequence substantially identical to the polypeptide sequences described herein, or fragments thereof.

A compound is "substantially pure" when it is separated from the components that naturally accompany it. Typically, a compound is substantially pure when it is at least 60%, more generally 75% or over 90%, by weight, of the total material in a sample. Thus, for example, a polypeptide that is chemically synthesized or produced by recombinant technology will be generally be substantially free from its naturally

associated components. A nucleic acid molecule may be substantially pure when it is not immediately contiguous with (i.e., covalently linked to) the coding sequences with which it is normally contiguous in the naturally occurring genome of the organism from which the DNA of the invention is derived. A nucleic acid molecule may also be substantially pure when it is isolated from the organism in which it is normally found. A substantially pure compound can be obtained, for example, by extraction from a natural source; by expression of a recombinant nucleic acid molecule encoding a polypeptide compound; or by chemical synthesis. Purity can be measured using any appropriate method such as column chromatography, gel electrophoresis, HPLC, etc.

A "substantially identical" sequence is an amino acid or nucleotide sequence that differs from a reference sequence only by one or more conservative substitutions, as discussed herein, or by one or more non-conservative substitutions, deletions, or insertions located at positions of the sequence that do not destroy the biological function of the amino acid or nucleic acid molecule. Such a sequence can be at least 10%, 20%, 30%, 40%, 50%, 52.5%, 55% or 60% or 75%, or more generally at least 80%, 85%, 90%, or 95%, or as much as 99% or 100% identical at the amino acid or nucleotide level to the sequence used for comparison using, for example, the Align Program (Myers and Miller, CABIOS, 1989, 4:11-17) or FASTA. For polypeptides, the length of comparison sequences may be at least 4, 5, 10, or 15 amino acids, or at least 20, 25, or 30 amino acids. In alternate embodiments, the length of comparison sequences may be at least 35, 40, or 50 amino acids, or over 60, 80, or 100 amino acids. For nucleic acid molecules, the length of comparison sequences may be at least 15, 20, or 25 nucleotides, or at least 30, 40, or 50 nucleotides. In alternate embodiments, the length of comparison sequences may be at least 60, 70, 80, or 90 nucleotides, or over 100, 200, or 500 nucleotides. Sequence identity can be readily measured using publicly available sequence analysis software (e.g., Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, Wis. 53705, or BLAST software available from the National Library of Medicine, or as described herein). Examples of useful software include the programs Pile-up and PrettyBox. Such software matches similar sequences by assigning degrees of homology to various substitutions, deletions, insertions, and other modifications.

Alternatively, or additionally, two nucleic acid sequences may be "substantially identical" if they hybridize under high stringency conditions. In some embodiments, high stringency conditions are, for example, conditions that allow hybridization comparable with the hybridization that occurs using a DNA probe of at least 500 nucleotides in length, in a buffer containing 0.5 M $NaHPO_4$, pH 7.2, 7% SDS, 1 mM EDTA, and 1% BSA (fraction V), at a temperature of 65°C, or a buffer containing 48% formamide, 4.8x SSC, 0.2 M Tris-Cl, pH 7.6, 1x Denhardt's solution, 10% dextran sulfate, and 0.1% SDS, at a temperature of 42°C. (These are typical conditions for high stringency northern or Southern hybridizations.) Hybridizations may be carried out over a period of about 20 to 30 minutes, or about 2 to 6 hours, or about 10 to 15 hours, or over 24 hours or more. High stringency hybridization is also relied upon for the success of numerous techniques routinely performed by molecular biologists, such as high stringency PCR, DNA sequencing, single strand conformational polymorphism analysis, and in situ hybridization. In contrast to northern and Southern hybridizations, these techniques are usually performed with relatively short probes (e.g., usually about 16 nucleotides or longer for PCR or sequencing and about 40 nucleotides or longer for in situ hybridization). The high stringency conditions used in these techniques are well known to those skilled in the art of molecular biology, and examples of them can be found, for example, in Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, New York, N.Y., 1998, which is hereby incorporated by reference.

The terms "nucleic acid" or "nucleic acid molecule" encompass both RNA (plus and minus strands) and DNA, including cDNA, genomic DNA, and synthetic (e.g., chemically synthesized) DNA. The nucleic acid may be double-stranded or single-stranded. Where single-stranded, the nucleic acid may be the sense strand or the antisense strand. A nucleic acid molecule may be any chain of two or more covalently bonded nucleotides, including naturally occurring or non-naturally occurring nucleotides, or nucleotide analogs or derivatives. By "RNA" is meant a sequence of two or more covalently bonded, naturally occurring or modified ribonucleotides. One example of a modified RNA included within this term is phosphorothioate RNA. By "DNA" is meant a sequence of two or more covalently bonded, naturally occurring or modified deoxyribonucleotides. By "cDNA" is meant complementary or copy DNA produced from an RNA template by the action of RNA-dependent DNA polymerase

(reverse transcriptase). Thus a "cDNA clone" means a duplex DNA sequence complementary to an RNA molecule of interest, carried in a cloning vector.

An "isolated nucleic acid" is a nucleic acid molecule that is free of the nucleic acid molecules that normally flank it in the genome or that is free of the organism in which it is normally found. Therefore, an "isolated" gene or nucleic acid molecule is in some cases intended to mean a gene or nucleic acid molecule which is not flanked by nucleic acid molecules which normally (in nature) flank the gene or nucleic acid molecule (such as in genomic sequences) and/or has been completely or partially purified from other transcribed sequences (as in a cDNA or RNA library). In some cases, an isolated nucleic acid molecule is intended to mean the genome of an organism such as a virus. An isolated nucleic acid of the invention may be substantially isolated with respect to the complex cellular milieu in which it naturally occurs. In some instances, the isolated material will form part of a composition (for example, a crude extract containing other substances), buffer system or reagent mix. In other circumstances, the material may be purified to essential homogeneity, for example as determined by PAGE or column chromatography such as HPLC. The term therefore includes, e.g., a genome; a recombinant nucleic acid incorporated into a vector, such as an autonomously replicating plasmid or virus; or into the genomic DNA of a prokaryote or eukaryote, or which exists as a separate molecule (e.g., a cDNA or a genomic DNA fragment produced by PCR or restriction endonuclease treatment) independent of other sequences. It also includes a recombinant nucleic acid which is part of a hybrid gene encoding additional polypeptide sequences. Preferably, an isolated nucleic acid comprises at least about 50, 80 or 90 percent (on a molar basis) of all macromolecular species present. Thus, an isolated gene or nucleic acid molecule can include a gene or nucleic acid molecule which is synthesized chemically or by recombinant means. Recombinant DNA contained in a vector are included in the definition of "isolated" as used herein. Also, isolated nucleic acid molecules include recombinant DNA molecules in heterologous host cells, as well as partially or substantially purified DNA molecules in solution. In vivo and in vitro RNA transcripts of the DNA molecules of the present invention are also encompassed by "isolated" nucleic acid molecules. Such isolated nucleic acid molecules are useful in the manufacture of the encoded polypeptide, as probes for isolating homologous sequences

(e.g., from other species), for gene mapping (e.g., by in situ hybridization with chromosomes), or for detecting expression of the nucleic acid molecule in tissue (e.g., human tissue, such as peripheral blood), such as by Northern blot analysis.

Various genes and nucleic acid sequences of the invention may be recombinant sequences. The term "recombinant" means that something has been recombined, so that when made in reference to a nucleic acid construct the term refers to a molecule that is comprised of nucleic acid sequences that are joined together or produced by means of molecular biological techniques. The term "recombinant" when made in reference to a protein or a polypeptide refers to a protein or polypeptide molecule which is expressed using a recombinant nucleic acid construct created by means of molecular biological techniques. The term "recombinant" when made in reference to genetic composition refers to a gamete or progeny with new combinations of alleles that did not occur in the parental genomes. Recombinant nucleic acid constructs may include a nucleotide sequence which is ligated to, or is manipulated to become ligated to, a nucleic acid sequence to which it is not ligated in nature, or to which it is ligated at a different location in nature. Referring to a nucleic acid construct as "'recombinant" therefore indicates that the nucleic acid molecule has been manipulated using genetic engineering, i.e. by human intervention. Recombinant nucleic acid constructs may for example be introduced into a host cell by transformation. Such recombinant nucleic acid constructs may include sequences derived from the same host cell species or from different host cell species, which have been isolated and reintroduced into cells of the host species. Recombinant nucleic acid construct sequences may become integrated into a host cell genome, either as a result of the original transformation of the host cells, or as the result of subsequent recombination and/or repair events.

As used herein, "heterologous" in reference to a nucleic acid or protein is a molecule that has been manipulated by human intervention so that it is located in a place other than the place in which it is naturally found. For example, a nucleic acid sequence from one species may be introduced into the genome of another species, or a nucleic acid sequence from one genomic locus may be moved to another genomic or extrachromasomal locus in the same species. A heterologous protein includes, for example, a protein expressed from a heterologous coding sequence or a protein

expressed from a recombinant gene in a cell that would not naturally express the protein.

By "antisense," as used herein in reference to nucleic acids, is meant a nucleic acid sequence that is complementary to one strand of a nucleic acid molecule. In some embodiments, an antisense sequence is complementary to the coding strand of a gene, preferably, a SARS virus gene. The preferred antisense nucleic acid molecule is one which is capable of lowering the level of polypeptide encoded by the complementary gene when both are expressed in a cell. In some embodiments, the polypeptide level is lowered by at least 10%, or at least 25%, or at least 50%, as compared to the polypeptide level in a cell expressing only the gene, and not the complementary antisense nucleic acid molecule.

A "probe" or "primer" is a single-stranded DNA or RNA molecule of defined sequence that can base pair to a second DNA or RNA molecule that contains a complementary sequence (the target). The stability of the resulting hybrid molecule depends upon the extent of the base pairing that occurs, and is affected by parameters such as the degree of complementarity between the probe and target molecule, and the degree of stringency of the hybridization conditions. The degree of hybridization stringency is affected by parameters such as the temperature, salt concentration, and concentration of organic molecules, such as formamide, and is determined by methods that are known to those skilled in the art. Probes or primers specific for SARS virus nucleic acid sequences or molecules may vary in length from at least 8 nucleotides to over 500 nucleotides, including any value in between, depending on the purpose for which, and conditions under which, the probe or primer is used. For example, a probe or primer may be 8, 10, 15, 20, or 25 nucleotides in length, or may be at least 30, 40, 50, or 60 nucleotides in length, or may be over 100, 200, 500, or 1000 nucleotides in length. Probes or primers specific for SARS virus nucleic acid molecules may have greater than 20-30% sequence identity, or at least 55-75% sequence identity, or at least 75-85% sequence identity, or at least 85-99% sequence identity, or 100% sequence identity to the nucleic acid sequences described herein. In various embodiments of the invention, probes having the sequences: 5'- ATg AAT TAC CAA gTC AAT ggT TAC –3', SEQ ID NO: 160; 5'- gAA gCT ATT CgT CAC gTT Cg-3', SEQ ID NO: 161; 5'- CTg TAg AAA ATC CTA gCT ggA g-3', SEQ ID NO: 162; 5'- CAT AAC CAg TCg

gTA CAg CTA-3', SEQ ID NO: 163; 5'- TTA TCA CCC gCgAAg AAg CT-3', SEQ ID NO: 164; 5'- CTC TAg TTg CATgAC AgC CCT C-3', SEQ ID NO: 165; 5'- TCg TgC gTg gAT TggCTT TgA TgT-3', SEQ ID NO: 166; 5'-ggg TTg ggA CTA TCC TAA gTg TgA-3', SEQ ID NO: 167; 5'-TAA CAC ACA AAC ACC ATC ATC A-3', SEQ ID NO: 168; 5'-ggT Tgg gAC TAT CCT AAg TgT gA-3', SEQ ID NO: 169; 5'-CCA TCA TCA gAT AgA ATC ATC ATA-3', SEQ ID NO: 170; 5'- CCT CTC TTg TTC TTg CTC gCA-3', SEQ ID NO: 171; 5'- TAT AgT gAg CCg CCA CAC Atg-3', SEQ ID NO: 172; 5'-TAACACACAACICCATCATCA-3', SEQ ID NO: 173; 5'-CTAACATGCTTAGGATAATGG-3', SEQ ID NO: 174; 5'-GCCTCTCTTGTTCTTGCTCGC-3', SEQ ID NO: 175; 5'-CAGGTAAGCGTAAAACTCATC-3', SEQ ID NO: 176; 5'-TACACACCTCAGCGTTG-3', SEQ ID NO: 177; 5'-CACGAACGTGACGAAT-3', SEQ ID NO: 178; 5'-GCCGGAGCTCTGCAGAATTC-3', SEQ ID NO: 179; 5'-CAGGAAACAGCTATGAC TTGCATCACCACTAGTTGTGCCACCAGGTT-3', SEQ ID NO: 180; 5'-TGTAAAACGACGGCCAGTTGATGGGATGGGACTATCCTAAGTGTGA-3', SEQ ID NO: 181; 5'- GCATAGGCAGTAGTTGCATC-3' , SEQ ID NO: 182, as well as sequences amplified by specific combinations of these probes, may be excluded from specific uses according to the invention. Probes can be detectably-labeled, either radioactively or non-radioactively, by methods that are known to those skilled in the art. Probes can be used for methods involving nucleic acid hybridization, such as nucleic acid sequencing, nucleic acid amplification by the polymerase chain reaction, single stranded conformational polymorphism (SSCP) analysis, restriction fragment polymorphism (RFLP) analysis, Southern hybridization, northern hybridization, in situ hybridization, electrophoretic mobility shift assay (EMSA), and other methods that are known to those skilled in the art.

By "complementary" is meant that two nucleic acid molecules, e.g., DNA or RNA, contain a sufficient number of nucleotides that are capable of forming Watson-Crick base pairs to produce a region of double-strandedness between the two nucleic acids. Thus, adenine in one strand of DNA or RNA pairs with thymine in an opposing complementary DNA strand or with uracil in an opposing complementary RNA strand. It will be understood that each nucleotide in a nucleic acid molecule need not form a

matched Watson-Crick base pair with a nucleotide in an opposing complementary strand to form a duplex.

By "vector" is meant a DNA molecule derived, e.g., from a plasmid, bacteriophage, or mammalian or insect virus, or artificial chromosome, that may be used to introduce a polypeptide, for example a SARS virus polypeptide, into a host cell by means of replication or expression of an operably linked heterologous nucleic acid molecule. By "operably linked" is meant that a nucleic acid molecule such as a gene and one or more regulatory sequences (e.g., promoters, ribosomal binding sites, terminators in prokaryotes; promoters, terminators, enhancers in eukaryotes; leader sequences, etc.) are connected in such a way as to permit the desired function e.g. gene expression when the appropriate molecules (e.g., transcriptional activator proteins) are bound to the regulatory sequences. A vector may contain one or more unique restriction sites and may be capable of autonomous replication in a defined host or vehicle organism such that the cloned sequence is reproducible. By "DNA expression vector" is meant any autonomous element capable of directing the synthesis of a recombinant peptide. Such DNA expression vectors include bacterial plasmids and phages and mammalian and insect plasmids and viruses. A "shuttle vector" is understood as meaning a vector which can be propagated in at least two different cell types, or organisms, for example vectors which are first propagated or replicated in prokaryotes in order for, for example, subsequent transfection into eukaryotic cells. A "replicon" is a unit that is capable of autonomous replication in a cell and may includes plasmids, chromosomes (e.g., mini-chromosomes), cosmids, viruses, etc. A replicon may be a vector.

A "host cell" is any cell, including a prokaryotic or eukaryotic cell, into which a replicon, such as a vector, has been introduced by for example transformation, transfection, or infection.

An "open reading frame" or "ORF" is a nucleic acid sequence that encodes a polypeptide. An ORF may include a coding sequence having i.e., a sequence that is capable of being transcribed into mRNA and/or translated into a protein when combined with the appropriate regulatory sequences. In general, a coding sequence includes a 5' translation start codon and a 3' translation stop codon.

A "leader sequence" is a relatively short nucleotide sequence located at the 5' end of an RNA molecule that acts as a primer for transcription.

A "transcriptional regulatory sequence" "TRS" or "intergenic sequence" is a nucleotide sequence that lies upstream of an open reading frame (ORF) and serves as a template for the reassociation of a nascent RNA strand-polymerase complex.

A "frameshift mutation" is caused by a shift in a open reading frame, generally due to a deletion or addition of at least one nucleotide, such that an alternative polypeptide is ultimately translated.

By "detectably labeled" is meant any means for marking and identifying the presence of a molecule, e.g., an oligonucleotide probe or primer, a gene or fragment thereof, a cDNA molecule, a polypeptide, or an antibody. Methods for detectably-labeling a molecule are well known in the art and include, without limitation, radioactive labeling (e.g., with an isotope such as ^{32}P or ^{35}S) and nonradioactive labeling such as, enzymatic labeling (for example, using horseradish peroxidase or alkaline phosphatase), chemiluminescent labeling, fluorescent labeling (for example, using fluorescein), bioluminescent labeling, antibody detection of a ligand attached to the probe, or detection of double-stranded nucleic acid. Also included in this definition is a molecule that is detectably labeled by an indirect means, for example, a molecule that is bound with a first moiety (such as biotin) that is, in turn, bound to a second moiety that may be observed or assayed (such as fluorescein-labeled streptavidin). Labels also include digoxigenin, luciferases, and aequorin.

A "peptide," "protein," "polyprotein" or "polypeptide" is any chain of two or more amino acids, including naturally occurring or non-naturally occurring amino acids or amino acid analogues, regardless of post-translational modification (e.g., glycosylation or phosphorylation). An "polyprotein", "polypeptide", "peptide" or "protein" of the invention may include peptides or proteins that have abnormal linkages, cross links and end caps, non-peptidyl bonds or alternative modifying groups. Such modified peptides are also within the scope of the invention. The term "modifying group" is intended to include structures that are directly attached to the peptidic structure (e.g., by covalent coupling), as well as those that are indirectly attached to the peptidic structure (e.g., by a stable non-covalent association or by covalent coupling to additional amino acid residues, or mimetics, analogues or

derivatives thereof, which may flank the core peptidic structure). For example, the modifying group can be coupled to the amino-terminus or carboxy-terminus of a peptidic structure, or to a peptidic or peptidomimetic region flanking the core domain. Alternatively, the modifying group can be coupled to a side chain of at least one amino acid residue of a peptidic structure, or to a peptidic or peptido-mimetic region flanking the core domain (e.g., through the epsilon amino group of a lysyl residue(s), through the carboxyl group of an aspartic acid residue(s) or a glutamic acid residue(s), through a hydroxy group of a tyrosyl residue(s), a serine residue(s) or a threonine residue(s) or other suitable reactive group on an amino acid side chain). Modifying groups covalently coupled to the peptidic structure can be attached by means and using methods well known in the art for linking chemical structures, including, for example, amide, alkylamino, carbamate or urea bonds.

A "polyprotein" is the polypeptide that is initially translated from the genome of a plus-stranded RNA virus, for example, a SARS virus. Accordingly, a polyprotein has not been subjected to post-translational processing by proteolytic cleavage into its processed protein products, and therefore, retains its cleavage sites. In some embodiments of the invention, the protease cleavage sites of a polyprotein may be modified, for example, by amino acid substitution, to result in a polyprotein that is incapable of being cleaved into its processed protein products.

An antibody "specifically binds" or "selectively binds" an antigen when it recognizes and binds the antigen, but does not substantially recognize and bind other molecules in a sample, having for example an affinity for the antigen which is 10, 100, 1000 or 10000 times greater than the affinity of the antibody for another reference molecule in a sample. A "neutralizing antibody" is an antibody that selectively interferes with any of the biological activities of a SARS virus polypeptide or polyprotein, for example, replication of the SARS virus, or infection of host cells. A neutralizing antibody may reduce the ability of a SARS virus polypeptide to carry out its specific biological activity by about 50%, or by about 70%, or by about 90% or more, or may completely abolish the ability of a SARS virus polypeptide to carry out its specific biological activity. Any standard assay for the biological activity of any SARS virus polypeptide, for example, assays determining expression levels, ability to infect host cells, or ability to replicate DNA, including those assays described herein or

known to those of skill in the art, may be used to assess potentially neutralizing antibodies that are specific for SARS virus polypeptides.

A "signal sequence" is a sequence of amino acids that may be identified, for example by homology or biological activity to a peptide sequence with the known function of targeting a polypeptide to a particular region of the cell. A signal sequence or signal peptide may be a peptide of any length, that is capable of targeting a polypeptide to a particular region of the cell. In some embodiments, the signal sequence may direct the polypeptide to the cellular membrane so that the polypeptide may be secreted. In alternate embodiments, the signal sequence may direct the polypeptide to an intracellular compartment or organelle, such as the Golgi apparatus, or to the surface of a virus, such as the SARS virus. In alternate embodiments, a signal sequence may range from about 13 or 15 amino acids in length to about 60 amino acids in length.

A "transmembrane protein" is an amphipathic protein having a hydrophobic region ("transmembrane domain") that spans the lipid bilayer of the cell membrane from the cytoplasm to the cell surface, or spans the viral envelope, interspersed between hydrophilic regions on both sides of the membrane. The number of hydrophobic regions in an amphipathic protein is often proportional to the number of times that proteins spans the lipid bilayer. Thus, a single transmembrane protein spans the lipid bilayer once, and has a single transmembrane domain, while a multi-transmembrane protein spans the lipid bilayer multiple times. Multi-transmembrane proteins may enable virus entry into a host cell, or act to initiate transduction of a signal from the cell surface to the interior of the cell, for example, by a conformational change upon ligand binding. A "transmembrane anchor" is a transmembrane domain that maintains a polypeptide in its position in the cell membrane or viral envelope and is generally hydrophobic. A transmembrane anchor may generally be in the structure of an alpha helix, i.e., a "transmembrane helix". Multi-transmembrane proteins may have multiple transmembrane alpha-helices.

A "nuclear localization signal" is an amino acid sequence that permits the entry of a polypeptide into the nucleus of a cell through nuclear pores. A nuclear localization signal generally has a cluster of positively charged residues, for example, lysines. A "lysine-rich sequence" is a sequence having at least two contiguous lysine residues, or

at least three contiguous lysine residues. In some embodiments, a lysine-rich sequence may be a nuclear localization signal.

An "ATP binding domain" is a consensus domain that is found in many ATP or GTP-binding proteins, and that forms a flexible loop (P-loop) between alpha-helical and beta pleated sheet domains. The general consensus for an ATP binding domain may be (A or G)-XXXXGK-(S or T).

A "RNA binding protein" is a protein that is capable of binding to a RNA molecule (see, for example, "RNA Binding Proteins: New Concepts in Gene Regulation" 1st ed, eds. K. Sandberg and S.E. Mulroney, Kluwers Academic Publishers, 2001). RNA binding proteins may contain common structural features such as arginine-rich tracts, for example, arginines alternating with aspartates, serines, or glycines, or zinc finger regions. RNA binding proteins may also have a common ribonucleotide sequence domain. RNA binding proteins are believed to play diverse roles in modulating post-transcriptional gene expression.

An "immune response" includes, but is not limited to, one or more of the following responses in a mammal: induction of antibodies, B cells, T cells (including helper T cells, suppressor T cells, cytotoxic T cells, γδ T cells) directed specifically to the antigen(s) in a composition or vaccine, following administration of the composition or vaccine. An immune response to a composition or vaccine thus generally includes the development in the host mammal of a cellular and/or antibody-mediated response to the composition or vaccine of interest. In general, the immune response will result in prevention or reduction of infection by a SARS virus.

An "immunogenic fragment" of a polypeptide or nucleic acid molecule refers to an amino acid or nucleotide sequence that elicits an immune response. Thus, an immunogenic fragment may include, without limitation, any portion of any of the SARS virus sequences described herein, or a sequence substantially identical thereto, that includes one or more epitopes (the antigenic determinant i.e., site recognized by a specific immune system cell, such as a T cell or a B cell). An "epitope" may include amino acids in a spatial orientation that they are non-contiguous in the amino acid sequence but are near each other due to the three dimensional conformation of the polypeptide. A epitope may include at least 3, 5, 8, or 10 or more amino acids. Immunogenic fragments or epitopes may be identified using standard methods known

to those of skill in the art, such as epitope mapping techniques or antigenicity or hydropathy plots using, for example, the Omiga version 1.0 program from Oxford Molecular Group (see, for example, U. S. Patent No. 4,708,871). Immunogenic fragments or epitopes may also be identified using methods for determining three dimensional molecule structure such as X-ray crystallography or nuclear magnetic resonance.

A "sample" may be a tissue biopsy, amniotic fluid, cell, blood, serum, plasma, urine, stool, sputum, conjunctiva, or any other specimen, or any extract thereof, obtained from a patient (human or animal), test subject, or experimental animal. A "sample" may also be a cell or cell line created under experimental conditions, and constituents thereof (such as cell culture supernatants, cell fractions, infected cells, etc.). The sample may be analyzed to detect the presence of a SARS virus gene, genome, polypeptide, nucleic acid molecule or virion, or to detect a mutation in a SARS virus gene, expression levels of a SARS virus gene or polypeptide, or the biological function of a SARS virus polypeptide, using methods that are known in the art. For example, methods such as sequencing, single-strand conformational polymorphism (SSCP) analysis, or restriction fragment length polymorphism (RFLP) analysis of PCR products derived from a sample can be used to detect a mutation in a SARS virus gene; ELISA or western blotting can be used to measure levels of SARS virus polypeptide or antibody affinity; northern blotting can be used to measure SARS mRNA levels, or PCR can be used to measure the level of a SARS virus nucleic acid molecule.

Other features and advantages of the invention will be apparent from the following description of the drawings and the invention, and from the claims.

Brief Description of the Drawings

Figures 1A-D show phylogenetic analyses of SARS proteins. Unrooted phylogenetic trees were generated by clustalw (Thompson, J. D. et al., *Nucleic Acids Res* 22, 4673-80, Nov 11, 1994) bootstrap analysis using 1000 iterations. Genbank accessions for protein sequences are as follows: Figure 1A: Replicase 1A: BoCov (Bovine Coronavirus):AAL40396, 229E (Human Coronavirus):NP_07355, MHV (Mouse Hepatitis Virus):NP_045298, AIBV (Avian Infectious bronchitis

virus):CAC39113, TGEV (Transmissible Gastroenteritis Virus): NP_058423. Figure 1B: Matrix Glycoprotein: PHEV (Porcine hemagglutinating encephalomyelitis virus):AAL80035, BoCov (Bovine Coronavirus):NP_150082, AIBV & AIBV2 (Avian infectious bronchitis virus): AAF35863 & AAK83027, MHV (Mouse hepatitis virus):AAF36439, TGEV (Transmissible gastroenteritis virus):NP_058427, 229E & OC43 (Human Coronavirus): NP_073555 & AAA45462, FCV (Feline coronavirus):BAC01160. Figure 1C: Nucleocapsid: MHV (Mouse hepatitis virus):P18446, BoCov (Bovine coronavirus):NP_150083, AIBV (Avian infectious bronchitis virus):AAK27162, FCV (Feline coronavirus):CAA74230, PTGV (Porcine transmissible gastroenteritis virus): AAM97563, 229E & OC43 (Human coronavirus):NP_073556 & P33469, PHEV (porcine hemagglutinating encephalomyelitis virus):AAL80036, TCV (Turkey coronavirus):AAF23873. Figure 1D: S (Spike) Protein: BoCov (Bovine coronavirus):AAL40400, MHV (Mouse hepatitis virus): P11225, OC43 & 229E (Human coronavirus):S44241 & AAK32191, PHEV (Porcine hemagglutinating encephalomyelitis virus):AAL80031, PRC (Porcine respiratory coronavirus):AAA46905, PEDV (Porcine epidemic diarrhea virus):CAA80971, CCov (Canine coronavirus):S41453, FICV (Feline infectious peritonitis virus):BAA06805, AIBV (Avian infectious bronchitis virus):AAO34396.

Figure 2 shows a schematic representation of the ORFs and s2m motif in the 29,736-base SARS virus genome.

Figures 3A-P show nucleotide sequences of the 29,736-base genome of the SARS virus (SEQ ID NOs: 1 and 2).

Figure 4 shows an alignment of the s2m regions from Avian infectious bronchitis virus (AIBV; SEQ ID NO: 32) and equine rhinovirus serotype 2 (ERV-2; SEQ ID NO: 31) with the 3' untranslated region (UTR; SEQ ID NO: 18) of the SARS virus (TOR2). The conserved areas in the s2m region are indicated by asterisks.

Figure 5 shows the amino acid sequence of the SARS virus S (Spike) Glycoprotein (SEQ ID NO: 33).

Figure 6 shows the amino acid sequence of the SARS virus M (Matrix) Glycoprotein (SEQ ID NO: 34).

Figure 7 shows the amino acid sequence of the SARS virus E (Small envelope) protein (SEQ ID NO: 35).

Figure 8 shows the amino acid sequence of the SARS virus N (Nucleocapsid) Protein (SEQ ID NO: 36).

Figure 9 shows an alignment of the matrix glycoprotein M from the SARS virus (Tor2_M or ORF5; SEQ ID NO: 34) and various other matrix glycoproteins (SEQ ID NOs: 37-43). Asterisks (*) indicate percentage identity to the SARS matrix protein as calculated by Align (Myers and Miller, CABIOS (1989) 4:11-17).

Figures 10A-B show an alignment of the nucleocapsid protein N from the SARS virus (Tor2_N; SEQ ID NO: 36) and various other nucleocapsid proteins (SEQ ID NOs: 44-52). Asterisks (*) indicate percentage identity to the SARS nucleocapsid protein calculated by Align (Myers and Miller, CABIOS (1989) 4:11-17) Figures 11A-K show the nucleotide sequence of the 29,751-base genome of the SARS virus (SEQ ID NO: 15).

Figure 12 shows a schematic representation of the ORFs and s2m motif in the 29,751-base SARS virus genome.

Figures 13A-D show phylogenetic analyses of SARS proteins. Unrooted phylogenetic trees were generated by clustalw 1.74 (J. D. Thompson, D. G. Higgins, T. J. Gibson, Nucleic Acids Res 22, 4673-80 (Nov 11, 1994) using the BLOSUM comparison matrix and a bootstrap analysis of 1000 iterations. Numbers indicate bootstrap replicates supporting each node. Phylogenetic trees were drawn with the Phylip Drawtree program 3.6a3 (Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle). Branch lengths indicate the number of substitutions per residue. Genbank accessions for protein sequences: A: Replicase 1A: BoCoV (Bovine Coronavirus):AAL40396, HCoV-229E (Human Coronavirus):NP_07355, MHV (Mouse Hepatitis Virus):NP_045298, IBV (Avian Infectious bronchitis virus):CAC39113, TGEV (Transmissible Gastroenteritis Virus): NP_058423. B: Membrane Glycoprotein: PHEV (Porcine hemagglutinating encephalomyelitis virus):AAL80035, BoCoV (Bovine Coronavirus):NP_150082, IBV & IBV2 (Avian infectious bronchitis virus): AAF35863 & AAK83027, MHV (Mouse hepatitis virus):AAF36439, TGEV (Transmissible gastroenteritis virus):NP_058427, HCoV-229E & HCoV-OC43 (Human Coronavirus): NP_073555 & AAA45462, FCoV (Feline coronavirus):BAC01160. C: Nucleocapsid: MHV (Mouse hepatitis virus):P18446,

BoCoV (Bovine coronavirus):NP_150083, IBV 1 & 2 (Avian infectious bronchitis virus): AAK27162 & NP_040838, FCoV (Feline coronavirus):CAA74230, PTGV (Porcine transmissible gastroenteritis virus): AAM97563, HCoV-229E & HCoV-OC43 (Human coronavirus):NP_073556 & P33469, PHEV (porcine hemagglutinating encephalomyelitis virus):AAL80036, TCV (Turkey coronavirus):AAF23873. D: S (Spike) Protein: BoCoV (Bovine coronavirus):AAL40400, MHV (Mouse hepatitis virus): P11225, HCoV-OC43 & HCoV-229E (Human coronavirus):S44241 & AAK32191, PHEV (Porcine hemagglutinating encephalomyelitis virus):AAL80031, PRCoV (Porcine respiratory coronavirus):AAA46905, PEDV (Porcine epidemic diarrhea virus):CAA80971, CCoV (Canine coronavirus):S41453, FIPV (Feline infectious peritonitis virus):BAA06805, IBV (Avian infectious bronchitis virus):AAO34396.

Figures 14A-F show an alignment of the spike glycoprotein S from the SARS virus (Tor2_S; SEQ ID NO: 33) and various other spike glycoproteins (SEQ ID NOs: 53-62). Asterisks (*) indicate percentage identity to the SARS spike protein as calculated by Align (Myers and Miller, CABIOS (1989) 4:11-17).

Figure 15 shows an alignment between the SARS virus Small envelope protein E (TOR2_E; SEQ ID NO: 35) and the Envelope protein (Protein 4) (X1 protein) (ORF 3) from Porcine transmissible gastroenteritis coronavirus (strain Purdue).Swissprot accession number P09048 (PGV; SEQ ID NO: 63), as calculated by FASTA (http://www.ebi.ac.uk/fasta33/).

Figures 16A-B show the amino acid sequence of the SARS virus Replicase 1A protein (SEQ ID NO: 64).

Figure 17 shows the amino acid sequence of the SARS virus Replicase 1B protein (SEQ ID NO: 65).

Figure 18 shows the amino acid sequence of ORF3 of SARS virus (SEQ ID NO: 66).

Figure 19 shows the amino acid sequence of ORF4 of SARS virus (SEQ ID NO: 67).

Figure 20 shows the amino acid sequence (SEQ ID NO: 68) of ORF6 (nucleotides 27059-27247 of the 29,736-base genome sequence) or ORF 7 (nucleotides 27,074-27,265 of the 29,751-base genome sequence) of SARS virus.

Figure 21 shows the amino acid sequence (SEQ ID NO: 69) of ORF7 (nucleotides 27258-27623 of the 29,736-base genome sequence) or ORF 8 (nucleotides 27,273-27,641 of the 29,751-base genome sequence),of SARS virus.

Figure 22 shows the amino acid sequence (SEQ ID NO: 70) of ORF8 (nucleotides 27623-27754 of the 29,736-base genome sequence) or ORF9 8 (nucleotides 27,638-27,772 of the 29,751-base genome sequence) of SARS virus.

Figure 23 shows the amino acid sequence (SEQ ID NO: 71) of ORF9 (nucleotides 27764-27880 of the 29,736-base genome sequence) or ORF10 (nucleotides 27,779-27,898 of the 29,751-base genome sequence) of SARS virus.

Figure 24 shows the amino acid sequence (SEQ ID NO: 72) of ORF10 (nucleotides 27849-28100 of the 29,736-base genome sequence) or ORF11 (nucleotides 27,864-28118 of the 29,751-base genome sequence) of SARS virus.

Figure 25 shows the amino acid sequence of ORF13 of SARS virus (SEQ ID NO: 73).

Figure 26 shows the amino acid sequence of ORF14 of SARS virus (SEQ ID NO: 74).

Figure 27 shows an alignment of the secreted region of the SARS virus ORF 10 of the 29,751-base genome sequence (sars) with the conotoxin from *Conus ventricosus* (conotoxin). Sequence identity is indicated by asterisks and sequence homology is indicated by dots.

Detailed Description of the Invention

In general, the invention provides nucleic acid molecules, polypeptides, and other reagents derived from a SARS virus, as well as methods of using such nucleic acid molecules, polypeptides, and other reagents.

The genome sequence (Figures 3A-P, 11A-K, SEQ ID NOs: 1, 2, and 15) reveals that the SARS coronavirus is only moderately related to other known coronaviruses, including two human coronaviruses, OC43 and 229E. Thus, the SARS virus is a previously unknown virus. The 5' end of the SARS genome contains a 5' leader sequence (Table 1; SEQ ID NO: 3) with sequence similarity to the highly conserved coronavirus core leader sequence, 5'-CUAAAC-3 (SEQ ID NO: 75;

Sawicki, S. G., et al., *Adv Exp Med Biol* 440, 215-9, 1998; Lai, M. M. and D. Cavanagh, *Adv Virus Res* 48, 1-100,1997). Transcriptional regulatory sequences (TRSs) were identified upstream of all open reading frames (ORFs) (Tables 1 and 2; SEQ ID NOs: 3-13 and 20-30). ORF9 and ORF10 of the 29,736-base SARS genome (ORF 10 and ORF 11 of the 29,751 base genome) overlap by 12 amino acids, and have matches to the TRS consensus in close proximity to their respective initiating methionine codons.

The 3' UTR sequence (SEQ ID NO: 18) of SARS virus contains a s2m region having the sequence ACATTTTCATCGAGGCCACGCGGAGTACGAT CGAGGGTACAGTGAAT; SEQ ID NO: 16) that includes a conserved, discontinuous 32 base-pair s2m motif. The conserved 32 base-pair motif is a universal feature of astroviruses that has also been identified in avian coronavirus (AIBV) and the ERV-2 equine rhinovirus. This motif has been identified by Jonassen C.M. et al. (J Gen Virol 1998 Apr;79 (Pt 4):715-8) as GCCGNGGCCACGC(G/C) GAGTA(C/G)GANCGAGGGTACAG(G/C) (SEQ ID NO: 19), where N is generally not part of the conserved motif, and can be any nucleotide. The region corresponding to the 32 base-pair motif in SARS virus includes the sequence: CGAGGCCACGCGGAGTACGATCGAGGGTACAG (SEQ ID NO: 17), and spans positions 29590-29621 of the 29,751 base genome. Figure 4 shows an alignment of the s2m regions from Avian infectious bronchitis virus (AIBV; SEQ ID NO: 32) and equine rhinovirus serotype 2 (ERV-2; SEQ ID NO: 31), as defined in Jonassen C.M. et al. (J Gen Virol 1998 Apr;79 (Pt 4):715-8), with the entire 3' untranslated region (UTR) of the SARS virus (TOR2) (SEQ ID NO: 18).

Table 1. Listing of the transcription regulatory sequences of the 29,736-base SARS genome, showing the nucleotide position (base) and associated open-reading frames (ORF). An asterisk (*) indicates consensus sequence.

Base	ORF	TRS Sequence	
45	Leader	TCTCTAAACGAACTTTAAAATCTGTG	(SEQ ID NO: 3)
21464	S	CAACTAAACGAACATG	(SEQ ID NO: 4)
25238	ORF3	CACATAAACGAACTTATG	(SEQ ID NO: 5)
26089	E	TGAGTACGAACTTATG	(SEQ ID NO: 6)
26326	M	GGTCTAAACGAACTAACT 40 ATG	(SEQ ID NO: 7)
26986	ORF6	AACTATAAATT 62 ATG	(SEQ ID NO: 8)
27244	ORF7	TCCATAAAACGAACATG	(SEQ ID NO: 9)
27575	ORF8	TGCTCTA---GTATTTTTAATACTTTG 24 ATG	(SEQ ID NO: 10)
27751	ORF9	AGTCTAAACGAACATG	(SEQ ID NO: 11)
27837	ORF10	CTAATAAACCTCATG	(SEQ ID NO: 12)
28084	N	TAAATAAACGAACAAATTAAAATG	(SEQ ID NO: 13)
		********	

Table 2. Listing of the transcription regulatory sequences of the 29,751-base SARS genome, showing the nucleotide position (base), associated open-reading frames (ORF), and identified transcription regulatory sequences. Numbers in parentheses within the alignment indicate distance to the putative initiating codon. The conserved core sequence is indicated in bold in the putative leader sequence. Contiguous sequences identical to region of the leader sequence containing the core sequence are shaded. No putative TRSs were detected for ORFs 4, 13 and 14, although ORF 13 may share the TRS associated with the N protein.

Base	ORF	TRS Sequence	
60	Leader	UCU**CUAAAC**GAACUUUAAAAUCUGUG	(SEQ ID NO: 20)
21479	S(Spike)	CAA**CUAAAC**GAACAUG	(SEQ ID NO: 21)
25252	ORF3	CACAU**AAACG**AACUUAUG	(SEQ ID NO: 22)
26104	Envelope	UGAGUA**CGAAC**UUAUG	(SEQ ID NO: 23)
26341	M	GGU**CUAAACG**AACUAACU (40)**AUG**	(SEQ ID NO:24)
27001	ORF7	AA**CUAAA**AAUU (62)**AUG**	(SEQ ID NO:25)
27259	ORF8	UCCAUAAAACGAACAUG	(SEQ ID NO: 26)
27590	ORF9	UGCUCUA---GUAUUUUUAAUACUUUG (24)**AUG**	(SEQ ID NO:27)
27766	ORF10	AGUCUAAACGAACAUG	(SEQ ID NO: 28)
27852	ORF11	CUAAUAAACCUCAUG	(SEQ ID NO: 29)
28099	NUCLEOCAPSID	UAAAUAAACGAACAAAUUAAAAUG	(SEQ ID NO: 30)

The coding potentials of the 29,736-base and 29,751-base genomes are depicted in Figures 2 and 12, respectively. Open reading frames (ORFs) include the Replicase 1a and 1b translation products, the Spike glycoprotein, the small Envelope protein, the Membrane and the Nucleocapsid protein. Construction of unrooted phylogenetic trees using this set of known proteins from representatives of the three known coronaviral groups reveals that the proteins encoded by the SARS virus do not readily cluster more closely with any known group than with any other (Figures 1A-D and 13A-D). In addition, nine novel ORFs have been analyzed.

The Replicase 1a ORF located at nucleotides 250-13395 of the 29,736-base genome, and nucleotides 265-13,398 of the 29,751-base genome, and replicase 1b ORF located at nucleotides 13395-21467 of the 29,736-base genome, and nucleotides 13,398 – 21,485 of the 29,751-base genome, occupy 21.2 kb of the SARS virus genome (Figures 2 and 12). These genes encode a number of proteins that are produced by proteolytic cleavage of a large polyprotein (Ziebuhr, J. et al., *J Gen Virol* 81, 853-79,

Apr, 2000). A frame shift mutation interrupts the protein-coding region, separating the 1a and 1b open-reading frames. The proteins encoded by the Replicase 1a and 1b ORFs are depicted in Figures 16A-B and 17, SEQ ID NOs: 64 and 65).

The Spike glycoprotein (S) (E2 glycoprotein gene; Figures 2 and 12; nucleotides 21477 to 25241 of the 29,736-base genome, and nucleotides 21,492 to 25,259 of the 29,751-base genome) encodes a surface projection glycoprotein precursor of about 1,255 amino acids in length (Figure 5; SEQ ID NO: 33), which may be significant in the virulence of the SARS virus. Mutations in this gene are correlated with altered pathogenesis and virulence in other coronaviruses (B. N. Fields et al., *Fields virology* (Lippincott Williams & Wilkins, Philadelphia, ed. 4th, 2001). In other coronaviruses, the mature spike protein is inserted in the viral envelope with the majority of the protein exposed on the surface of the particles. Three molecules of the Spike protein form the characteristic peplomers or corona-like structures of this virus family. Analysis of the spike glycoprotein with SignalP (Nielson, H. et al., *Prot Engineer. 10*:1-6 (1997) indicates a signal peptide (MFIFLLFLTLTSG; SEQ ID NO: 76)(probability 0.996) with cleavage between residues 13 and 14. TMHMM (Sonnhammer, E. L. et al., *Proc Int Conf Intell Syst Mol Biol 6,* 175-82 (1998)) indicates a transmembrane domain near the C-terminal end (WYVWLGFIAGLIAIVMVTILLCC; SEQ ID NO: 183). Together these data indicate a type I membrane protein with N-terminus and the majority of the protein (residues 14-1195) on the outside of the cell-surface or virus particle, which may be responsible for binding to a cellular receptor. The SARS virus Spike glycoprotein has limited sequence identity to other, known Spike glycoproteins (Figures 14A-F).

ORF 3 (Figures 2 and 12; nucleotides 25253-26074 of the 29,736-base genome and nucleotides 25,268 - 26,092 of the 29,751-base genome) encodes a protein of 274 amino acids (Figure 18; SEQ ID NO: 66) that lacks significant similarities to any known protein when analyzed with BLAST (Altschul, S. F. et al., *Nucleic Acids Res* 25, 3389-402, Sep 1, 1997), FASTA (Pearson, W. R. and D. J. Lipman, *Proc Natl Acad Sci USA* 85, 2444-8, Apr, 1988) or PFAM (Bateman, A. et al., *Nucleic Acids Res* 30, 276-80, Jan 1, 2002). Analysis of the N-terminal 70 amino acids with SignalP indicates the existence of a signal peptide (MDLFMRFFTLRSITAQ; SEQ ID NO: 184) and a cleavage site (probability 0.540). Both TMpred (Hofman, K. and W. Stoffel, *Biol.*

Chem. Hoope-Seyler 374, 166 (1993) and TMHMM indicate three trans-membrane regions spanning approximately residues 34-56 (TIPLQASLPFGWLVIGVAFLAVF, SEQ ID NO: 77), 77-99 (FQFICNLLLLFVTIYSHLLLVAA, SEQ ID NO: 78), and 103-125 (AQFLYLYALIYFLQCINACRIIM, SEQ ID NO: 79). Both TMpred and TMHMM indicate that the C-terminus and a large 149 amino acid domain is located inside the viral or cellular membrane. The C-terminal (interior) region of the protein, corresponding to about amino acids 124-274 (MRCWLCWKCKSKNPLLYDANYFVCWHTHNYDYCIPYNSVTDTIVVTEGDGI STPKLKEDYQIGGYSEDRHSGVKDYVVVHGYFTEVYYQLESTQITTDTGIENAT FFIFNKLVKDPPNVQIHTIDGSSGVANPAMDPIYDEPTTTTSVPL; SEQ ID NO: 185) may encode a protein domain with ATP-binding properties (PD037277).

ORF 4 (Figure 12; nucleotides 25,689 - 26,153 of the 29,751-base genome) encodes a predicted protein of 154 amino acids (Figure 19; SEQ ID NO: 67). This ORF overlaps entirely with ORF 3 and the E protein. ORF4 may be expressed from the ORF mRNA using an internal ribosomal entry site. BLAST analyses failed to identify matching sequences. Analysis with TMPred predicts a single transmembrane helix, amino acids 1-20 MMPTTLFAGTHITMTTVYHI, SEQ ID NO: 186.

The small envelope protein E (Figures 2 and 12; nucleotides 26102-26329 of the 29,736-base genome and nucleotides 26,117 - 26,347, ORF 5, of the 29,751-genome) encodes a protein of 76 amino acids (Figure 7; SEQ ID NO: 35). BLAST and FASTA comparisons indicate that the protein, while novel, is homologous to multiple envelope proteins (alternatively known as small membrane proteins) from several coronaviruses. An alignment of the SARS virus E protein with the envelope protein of Porcine transmissible gastroenteritis coronavirus indicates approximately 28% sequence identity between the two proteins over a 61 amino acid overlap, as calculated by FASTA (Figure 15). PFAM analysis of the protein indicates that the small envelope protein E is a member of the NS3_EnvE protein family. InterProScan (R. Apweiler et al., *Nucleic Acids Res* 29, 37-40, Jan 1, 2001; Zdobnov, E. M. and R. Apweiler, *Bioinformatics* 17, 847-8, Sep, 2001) analysis indicates that the protein is a component of the viral envelope, and homologs of it are found in other viruses, including gastroenteritis virus and murine hepatitis virus. SignalP analysis indicates the presence of a transmembrane anchor (probability 0.939). TMpred analysis indicates a similar

transmembrane anchor at positions 17-34 (VLLFLAFVVFLLVTLAIL, SEQ ID NO: 80), which is consistent with the known association of homologous proteins with the viral envelope. TMHMM indicates a type II membrane protein with the majority of the 46 residue C terminus hydrophilic domain (TALRLCAYCCNIVNVSLVKPTVYVYSRVKNLNSSEGVPDLLV; SEQ ID NO: 187) located on the surface of the viral particle. The E protein may be important for viral replication.

The Matrix glycoprotein M (Figures 2 and 12; nucleotides 26383-27045 of the 29,736-base genome and nucleotides 26,398 - 27,063, ORF 6, of the 29,751-genome) encodes a protein of 221 amino acids (Figure 6; SEQ ID NO: 34). BLAST and FASTA analysis of the protein, while novel, reveals homologies to coronaviral matrix glycoproteins (Figure 9). The association of the spike glycoprotein (S) with the matrix glycoprotein (M) may be an essential step in the formation of the viral envelope and in the accumulation of both proteins at the site of virus assembly. Analysis of the amino acid sequence with SignalP indicates a signal sequence (probability 0.932), located at approximately residues 1-39 (MADNGTITVEELKQLLEQWNLVIGFLFLAWIMLLQFAYS; SEQ ID NO: 188) that is unlikely to be cleaved. TMHMM and TMpred analysis both indicate the presence of three trans-membrane helices, located at approximately residues 15-37 (LLEQWNLVIGFLFLAWIMLLQFA; SEQ ID NO: 81), 50-72 (LVFLWLLWPVTLACFVLAAVYRI; SEQ ID NO: 82) and 77-99 (GGIAIAMACIVGLMWLSYFVASF; SEQ ID NO: 83), with the 121 amino acid hydrophilic domain on the inside of the virus particle, where it may interact with nucleocapsid. The hydrophilic domain may run from approximately amino acids PLRGTIVTRPLMESELVIGAVIIRGHLRMAGHSLGRCDIKDLPKEITVATSRTLS YYKLGASQRVGTDSGFAAYNRYRIGNYKLNTDHAGSNDNIALLVQ (SEQ ID NO: 189) i.e. approximately amino acids 95 or 99 to 221 of SEQ ID NO: 34. PFAM analysis reveals a match to PFAM domain PF01635, and alignments to 85 other sequences in the PFAM database bearing this domain, which is indicative of the coronavirus matrix glycoprotein.

ORF6 (Figure 2; nucleotides 27059-27247 of the 29,736-base genome sequence) or ORF 7 (Figure 12; nucleotides 27,074-27,265 of the 29,751-base genome

sequence) encodes a protein of 63 amino acids (Figure 20; SEQ ID NO: 68). TMpred analysis indicates a trans-membrane helix located between residues 3 or 4 and 22 (HLVDFQVTIAEILIIIMRTF; SEQ ID NO: 84), with the N-terminus located outside the viral particle.

Similarly, the gene encoding ORF7 (Figure 2; nucleotides 27258-27623 of the 29,736-base genome sequence) or ORF 8 (Figure 12; nucleotides 27,273-27,641 of the 29,751-base genome sequence), encoding a protein of 122 amino acids (Figure 21; SEQ ID NO: 69), has no significant BLAST or FASTA matches to known proteins. Analysis of this sequence with SignalP indicates a cleaved signal sequence (MKIILFLTLIVFTSC; SEQ ID NO: 85) (probability 0.995), with the cleavage site located between residues 15 and 16. TMpred and TMHMM analysis also indicates a trans-membrane helix located approximately at residues 99-117 (SPLFLIVAALVFLILCFTI; SEQ ID NO: 86). Together these data indicate that this protein is a type I membrane protein with the major hydrophilic domain of the protein (residues 16-98; ELYHYQECVRGTTVLLKEPCP SGTYEGNSPFHPLADNKFALTCTSTHFAFACADGTRHTYQLRARSVSPKLFIRQ EEVQQELY; SEQ ID NO: 87) and the amino-terminus is oriented inside the lumen of the ER/Golgi, or on the surface of the cell membrane or virus particle,depending on the membrane localization of the protein.

ORF8 (Figure 2; nucleotides 27623-27754 of the 29,736-base genome sequence) or ORF9 (Figure 12; nucleotides 27,638-27,772 of the 29,751-base genome sequence), encodes a protein of 44 amino acids (Figure 22; SEQ ID NO: 70). FASTA analysis of this sequence revealed some weak similarities (37% identity over a 35 amino acid overlap) to Swiss-Prot accession Q9M883, annotated as a putative sterol-C5 desaturase. A similarly weak match to a hypothetical *Clostridium perfringens* protein (Swiss-Prot accession CPE2366) was also detected. TMpred indicated a single strong trans-membrane helix FYLCFLAFLLFLVLIMLIIFWFS, SEQ ID NO: 190, with little preference for alternate models in which the N-terminus was located inside or outside the particle.

Similarly ORF9 (Figure 2; nucleotides 27764-27880 of the 29,736-base genome sequence) or ORF10 (Figure 12; nucleotides 27,779-27,898 of the 29,751-base genome sequence) encoding a protein of 39 amino acids (Figure 23; SEQ ID NO: 71), exhibited

no significant matches in BLAST and FASTA searches but encodes a trans-membrane helix LLIVLTCISLCSCICTVVQ (SEQ ID NO: 191) by TMPred, with the N-terminus located within the viral particle. The region immediately upstream of this protein exhibits a strong match to the TRS consensus (Table 2), indicating that a transcript initiates from this site. The large number of cysteine residues (6) may result in cross linking of the amino acids. Amino acids ICTVVQRCASNKPHVLEDPCKVQH (SEQ ID NO: 192) of this protein may be secreted. The secreted amino acids exhibit homology to toxin proteins, for example, to the conotoxin of *Conus ventricosus* (Figure 27). Antigenic peptides from the hydrophilic (secreted) region, for example, CICTVVQRCASNKPHVLEDPCK (SEQ ID NO: 193), were used to generate monoclonal antibodies using standard techniques. Furthermore, the C terminal amino acids form a sequence that shares homology to farnesylation sites (CKQH), which generally require C terminal location to be functional. This protein may act as a virulence factor and/or may facilitate transmission to humans.

ORF10 (Figure 2; nucleotides 27849-28100 of the 29,736-base genome sequence) or ORF11 (Figure 12; nucleotides 27,864-28118 of the 29,751-base genome sequence) encoding a protein of 84 amino acids (Figure 24; SEQ ID NO: 72) exhibited only very short (9-10 residues) matches to a region of the human coronavirus E2 glycoprotein precursor (starting at residue 801). Analysis by SignalP and TMHMM predict a soluble protein. A detectable alignment to the TRS consensus sequence was also found (Table 2).

The protein (422 amino acids; Figure 8; SEQ ID NO: 36) encoded by the Nucleocapsid gene (Figure 2; nucleotides 28105-29370 of the 29,736-base genome sequence; Figure 12, nucleotides 28,120-29,388 of the 29,751-base genome sequence) aligns well with nucleocapsid proteins from other representative coronaviruses (Figures 10A-B), although a short lysine rich region (KTFPPTEPKKDKKKKTDEAQ; SEQ ID NO: 14) is unique to SARS. This region is suggestive of a nuclear localization signal Since some coronaviruses are able to replicate in enucleated cells, the SARS virus nucleocapsid protein may have evolved a novel nuclear function, which may play a role in pathogenesis. In addition, the basic nature of this peptide suggests it may assist in RNA binding. The SARS nucleocapsid protein is also a good candidate for diagnostic tests.

ORF 13 (Fig. 12; nucleotides 28,130 – 28,426 of the 29,751-base genome sequence) encodes a novel protein of 98 amino acids (Figure 25; SEQ ID NO: 73). ORF 14 (Fig. 12; nucleotides 28,583 – 28,795 of the 29,751-base genome sequence) encodes a novel protein of 70 amino acids (Figure 26; SEQ ID NO: 74). TMPred predicts a single transmembrane helix VVAVIQEIQLLAAVGEILLLEW (SEQ ID NO: 194).

Various features of the SARS virus genome are summarised in Table 3. While Table 3 refers to the 29,751-base genome sequence, the features are also applicable to the 29,736-base genome sequence (SEQ ID NOs: 1 and 2).

Table 3. Features of the SARS virus 29,751-base genome sequence.

Feature	Start – End[1]	No. amino acids	No. bases	Frame	TRS
Orf 1a	265 - 13,398	4,382	13,149	+1	N/A
Orf 1b	13,398 – 21,485	2,628	7,887	+3	N/A
S protein	21,492 – 25,259	1,255	3,768	+3	Strong
Orf 3	25,268 – 26,092	274	825	+2	Strong
Orf 4	25,689 – 26,153	154	465	+3	Absent[2]
E protein	26,117 – 26,347	76	231	+2	Weak
M protein	26,398 – 27,063	221	666	+1	Strong
Orf 7	27,074 – 27,265	63	192	+2	Weak
Orf 8	27,273 – 27,641	122	369	+3	Strong
Orf 9	27,638 – 27,772	44	135	+2	Weak
Orf 10	27,779 – 27,898	39	120	+2	Strong
Orf 11	27,864 – 28,118	84	255	+3	Weak
N protein	28,120 – 29,388	422	1,269	+1	Strong
Orf 13[3]	28,130 – 28,426	98	297	+2	Absent[2]
Orf 14[3]	28,583 – 28,795	70	213	+2	Absent
s2m motif	29,590 – 29,621	N/A	30	N/A	N/A

1. End coordinates include the stop codon, except for ORF 1a and s2m.
2 These ORFs overlap substantially or completely with other and may share TRSs.
N/A indicates not applicable.

Various polymorphisms may exist in the SARS virus. In the SARS 29,736-base genome sequences (SEQ ID NO: 1 or 2), for example, nucleotides 7904, 16607, 19168,

24857, or 26842 may be C or T; or nucleotides 19049, 23205, or 25283 may be G or A, and in the SARS 29,751-base genome sequence (SEQ ID NO: 15), for example, nucleotides 7919, 16622, 19183, 24872, or 26857 may be C or T; or nucleotides 19064, 23220, or 25298 may be G or A. In some embodiments, the nucleotide changes may result in no change in the encoded amino acid, or in a conservative or non-conservative change in the encoded amino acid. In some embodiments, a nucleotide change, as described herein, at position 7904 or 7919, may result in a A to V amino acid substitution, in the Replicase 1A protein coding region; a change at position 19168 or 19183 may result in a V to A amino acid substitution, in the Replicase IB protein coding region; a change at position 23205 or 23220 may result in a A to S amino acid substitution (non-conservative change), affecting the Spike glycoprotein coding region; a change at position 25283 or 25298 may result in a R to G amino acid substitution (non-conservative change), affecting ORF3; or a change at position 26842 or 26857 may result in a S to P amino acid substitution (non-conservative change), affecting the Nucleocapsid protein coding region, in the SARS 29,736-base (SEQ ID NO: 1 or 2) and 29,751-base genome (SEQ ID NO: 15) sequences, respectively. In various embodiments, a nucleotide or amino acid sequence including a particular polymorphism may be selected, for example, for use in the methods of the invention, or may be excluded, for example, from a particular use according to the invention.

Various alternative embodiments of the invention are described below. These embodiments include, without limitation, identification and use of SARS virus nucleic acid and amino acid sequences for diagnostic or therapeutic uses.

Diagnosis of SARS virus-related disorders

A SARS virus-related disorder is any disorder that is mediated by the SARS virus, or by a nucleic acid molecule or polypeptide derived from the SARS virus. Accordingly, SARS virus nucleic acid molecules and polypeptides may be used to diagnose and identify a SARS virus-related disorder in a mammal, for example, a human or a domestic, farm, wild, or experimental animal. In some embodiments, SARS virus nucleic acid molecules and polypeptides may be used to screen such animals, e.g., civet cats, for the presence of SARS virus. A SARS virus-related disorder may be a hepatic, enteric, respiratory, or neurological disorder, and may be

accompanied by one or more symptoms or indications including, but not limited to, fever, cough, shortness of breath, headache, low blood oxygen concentration, liver damage, or reduced lymphocyte numbers. Accordingly, samples for diagnosis may be obtained from cells, blood, serum, plasma, urine, stool, conjunctiva, sputum, asopharyngeal or oropharyngeal swabs, tracheal aspirates, bronchalveolar lavage, pleural fluid, amniotic fluid, or any other specimen, or any extract thereof, or by tissue biopsy of for example lungs or major organs, obtained from a patient (human or animal), test subject, or experimental animal.

A SARS virus-related disorder may be diagnosed by amplifying a SARS nucleic acid molecule or fragment thereof from a sample. Probes or primers for use in amplification may be prepared using standard techniques. In some embodiments, probes or primers are selected from regions of a SARS virus genome as described herein that show limited sequence homology or identity (e.g., less than 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, or 100% identity) to other viruses or pathogens, or to host sequences.

Nucleic acid sequences can be amplified as needed by methods known in the art. For example, this can be accomplished by e.g., polymerase chain reaction "PCR" of DNA or of RNA by reverse transcriptase-PCR or "RT-PCR" (See generally PCR Technology: Principles and Applications for DNA Amplification (ed. H. A. Erlich, Freeman Press, NY, N.Y., 1992); PCR Protocols: A Guide to Methods and Applications (eds. Innis, et al., Academic Press, San Diego, Calif., 1990); Mattila et al., Nucleic Acids Res. 19, 4967 (1991); Eckert et al., PCR Methods and Applications 1, 17 (1991); PCR (eds. McPherson et al., IRL Press, Oxford); and U.S. Pat. No. 4,683,202 issued July 28, 1987 to Mullis) Variations of standard PCR techniques, such as for example real time RT-PCR using internal as well as amplification primers, resulting in increased sensitivity and speed, and reduction of risk of sample contamination (see for example Higuchi, R., et al., "Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions," Bio/Technology, vol. 11, pp. 1026-1030 (1993); Heid et al, "Real Time Quantitative PCT", Genome Research, 1996, pp. 986-994; Gibson UE et al., "A novel method for real time quantitative RT-PCR," Genome Res. 1996 Oct;6(10):995-1001), or the "Tacman" approach to PCR, described by for example Holland et al, Proc. Natl. Acad. Sci., 88: 7276-7280 (1991), may be performed.

Other suitable amplification and analytical methods include the single base primer extension (see for example U.S. Patent No. 6,004,744), mini-sequencing, ligase chain reaction (LCR) (see for example Wu and Wallace, Genomics 4, 560 (1989), Landegren et al., Science 241, 1077 (1988), transcription amplification (Kwoh et al., Proc. Natl. Acad. Sci. USA 86, 1173 (1989)), and self-sustained sequence replication (Guatelli et al., Proc. Nat. Acad. Sci. USA, 87, 1874 (1990)) and nucleic acid based sequence amplification (NASBA). The latter two amplification methods involve isothermal reactions based on isothermal transcription, which produce both single stranded RNA (ssRNA) and double stranded DNA (dsDNA) as the amplification products in a ratio of about 30 or 100 to 1, respectively.

A SARS virus-related disorder may also be diagnosed using an antibody directed against a SARS virus nucleic acid or amino acid sequence that specifically binds a nucleic acid molecule or polypeptide. In an alternative embodiment, the antibody may be directed against a SARS polypeptide, for example, the S polypeptide or fragment thereof that is located on the surface of the SARS virion. Methods for preparation of antibodies or for assaying antibody binding are well known in the art.

Serological diagnosis may included detection of antibodies against a SARS virus polypeptide or nucleic acid molecule, e.g., the Nucleocapsid protein, produced in response to infection using techniques such as indirect fluorescent antibody testing or enzyme-linked immunosorbent assays (ELISA). A SARS virus-related disorder may also be diagnosed by for example performing *in situ* probe hybridization studies on tissue specimens.

In some aspects, diagnostic tests as described herein or known to those of skill in the art may be performed for SARS virus variants that exhibit increased pathogenicity, such as strains having redundant sequences.

In some embodiments, reagents for diagnosis (e.g, probes, primers, antibodies, etc.) may be provided in kits which may optionally include instructions for using the reagent or may include other reagents for performing the appropriate assay e.g., controls, standards, buffers, etc.

Therapy or Prophylaxis for SARS virus-related disorders

Compounds according to the invention may also be used to provide therapeutics or prophylactics for SARS virus-related disorders. Accordingly, such compounds may be used to treat a mammal, for example, a human or a domestic, farm, wild, or experimental animal that has or is at risk for a SARS virus-related disorder. Such compounds may include, without limitation, compounds that interfere with SARS virus replication, expression of SARS virus proteins, or the ability of the SARS virus to infect a host cell. Accordingly, in some embodiments, compounds that act as antagonists to SARS virus polypeptides may be used as therapeutics or prophylactics for SARS virus related disorders. In some embodiments, purified SARS virus polypeptides may be used as for example competitive inhibitors to disrupt viral function. For example, a Spike protein lacking a functional domain, or having some other modification that maintains binding but reduces or eliminates pathogenicity, may be used to disrupt viral function. In some embodiments, antibodies that bind SARS virus polypeptides or nucleic acid molecules, for example, humanized antibodies, may be used as therapeutics or prophylactics.

In some embodiments, the SARS-virus compounds may be used as vaccines, or may be used to develop vaccines. For example, peptides derived from portions of SARS-virus proteins or polypeptides located on the outside of the virion or cell surface may be useful for vaccines or for generation of therapeutic or prophylactic antibodies.

A "vaccine" is a composition that includes materials that elicit a desired immune response. A vaccine may select, activate or expand memory B and T cells of the immune system to, for example, enable the elimination of infectious agents, such as a SARS virus, or a component thereof. In some embodiments, a vaccine includes a suitable carrier, such as an adjuvant, which is an agent that acts in a non-specific manner to increase the immune response to a specific antigen, or to a group of antigens, enabling the reduction of the quantity of antigen in any given vaccine dose, or the reduction of the frequency of dosage required to generate the desired immune response.

Vaccines according to the invention may include SARS virus polypeptides and nucleic acid molecules described herein, or immunogenic fragments thereof. In some embodiments, a SARS virus Spike polypeptide, Envelope polypeptide, or membrane glycoprotein or fragments thereof may be suitable for vaccine applications. In some

embodiments, the vaccines may be multivalent and include one or more epitopes from a SARS virus polypeptide or fragment thereof.

In some embodiments of the invention, a vaccine may include a live or killed microorganism e.g., a SARS virus or a component thereof. If a live SARS virus is used, which may be administered in the form of an oral vaccine, is may contain non-revertible genetic alterations (for example, large deletions or insertions in the genomic sequence) that reduce or eliminate the virulence of the virus ("attenuated virus"), but not its induction of an immune response. In some embodiments, a live vaccine may include an attenuated non-SARS microorganism (e.g, bacteria or virus such as vaccinia virus) that is capable of expressing a SARS virus polypeptide or immunogenic fragment thereof as described herein. In some embodiments, a vaccine may include SARS virus polypeptides or nucleic acid molecules having modifications that facilitate ease of administration. For example, an indigestible SARS virus polypeptide or nucleic acid molecule may be used for oral administration, and a modification that is suitable for inhalation may be used for administration to the lung.

A "nucleic acid vaccine" or "DNA vaccine" as used herein, is a nucleic acid construct comprising a polynucleotide encoding a polypeptide antigen, particularly an antigenic amino acid subsequence identified by methods described herein or known in the art. The nucleic acid construct can also include transcriptional promoter elements, enhancer elements, splicing signals, termination and polyadenylation signals, and other nucleic acid sequences. Thus, a nucleic acid vaccine is generally introduced into a subject animal using for example one or more DNA plasmids including one or more antigen-coding sequences (for example, a SARS virus Envelope polypeptide or membrane glycoprotein sequence) that are capable of transfecting cells in vivo and inducing an immune response (see for example Whalen RG et al. DNA-mediated immunization and the energetic immune response to hepatitis B surface antigen. Clin Immunol Immunopathol 1995;75:1-12; Wolff JA et al. Direct gene transfer into mouse muscle in vivo. Science 1990;247:1465-8; Fynan EF et al. DNA vaccines: protective immunizations by parental, mucosal, and genegun inoculations. Proc Natl Acad Sci USA 1993; 90:11478-82). In some embodiments, a library of nucleic acid fragments may be prepared by cloning SARS virus genomic DNA into a plasmid expression vector using known techniques and the library then used as a nucleic acid vaccine (see

for example Barry MA, et al. Protection against mycoplasma infection using expression-library immunization. Nature 1995;377:632-5).

The subject is administered the nucleic acid vaccine using standard methods. The vertebrate can be administered parenterally, subcutaneously, intravenously, intraperitoneally, intradermally, intramuscularly, topically, orally, rectally, nasally, buccally, vaginally, by inhalation spray, or via an implanted reservoir in dosage formulations containing conventional non-toxic, physiologically acceptable carriers or vehicles. Alternatively, the subject is administered the nucleic acid vaccine through the use of a particle acceleration or bombardment instrument (a "gene gun"). The form in which it is administered (e.g., capsule, tablet, solution, emulsion) will depend in part on the route by which it is administered. For example, for mucosal administration, nose drops, inhalants or suppositories can be used. The nucleic acid vaccine can be administered in conjunction with known adjuvants. The adjuvant is administered in a sufficient amount, which is that amount that is sufficient to generate an enhanced immune response to the nucleic acid vaccine. The adjuvant can be administered prior to (e.g., 1 or more days before) inoculation with the nucleic acid vaccine; concurrently with (e.g., within 24 hours of) inoculation with the nucleic acid vaccine; contemporaneously (simultaneously) with the nucleic acid vaccine (e.g., the adjuvant is mixed with the nucleic acid vaccine, and the mixture is administered to the vertebrate); or after (e.g., 1 or more days after) inoculation with the nucleic acid vaccine. The adjuvant can also be administered at more than one time (e.g., prior to inoculation with the nucleic acid vaccine and also after inoculation with the nucleic acid vaccine). As used herein, the term "in conjunction with" encompasses any time period, including those specifically described herein and combinations of the time periods specifically described herein, during which the adjuvant can be administered so as to generate an enhanced immune response to the nucleic acid vaccine (e.g., an increased antibody titer to the antigen encoded by the nucleic acid vaccine, or an increased antibody titer to the pathogenic agent). The adjuvant and the nucleic acid vaccine can be administered at approximately the same location on the vertebrate; for example, both the adjuvant and the nucleic acid vaccine are administered at a marked site on a limb of the subject.

In some embodiments, expression of a SARS virus gene or coding or non-coding region of interest may be inhibited or prevented using RNA interference (RNAi)

technology, a type of post-transcriptional gene silencing. RNAi may be used to create a functional "knockout", i.e. a system in which the expression of a gene or coding or non-coding region of interest is reduced, resulting in an overall reduction of the encoded product. As such, RNAi may be performed to target a nucleic acid of interest or fragment or variant thereof, to in turn reduce its expression and the level of activity of the product which it encodes. Such a system may be used for therapy or prophylaxis, as well as for functional studies. RNAi is described in for example published US patent applications 20020173478 (Gewirtz; published November 21, 2002) and 20020132788 (Lewis *et al.*; published November 7, 2002). Reagents and kits for performing RNAi are available commercially from for example Ambion Inc. (Austin, TX, USA) and New England Biolabs Inc. (Beverly, MA, USA).

The initial agent for RNAi in some systems is thought to be dsRNA molecule corresponding to a target nucleic acid. The dsRNA is then thought to be cleaved into short interfering RNAs (siRNAs) which are 21-23 nucleotides in length (19-21 bp duplexes, each with 2 nucleotide 3' overhangs). The enzyme thought to effect this first cleavage step has been referred to as "Dicer" and is categorized as a member of the Rnase III family of dsRNA-specific ribonucleases. Alternatively, RNAi may be effected via directly introducing into the cell, or generating within the cell by introducing into the cell a suitable precursor (e.g. vector, etc.) of such an siRNA or siRNA-like molecule. An siRNA may then associate with other intracellular components to form an RNA-induced silencing complex (RISC). The RISC thus formed may subsequently target a transcript of interest via base-pairing interactions between its siRNA component and the target transcript by virtue of homology, resulting in the cleavage of the target transcript approximately 12 nucleotides from the 3' end of the siRNA. Thus the target mRNA is cleaved and the level of protein product it encodes is reduced.

RNAi may be effected by the introduction of suitable *in vitro* synthesized siRNA or siRNA-like molecules into cells. RNAi may for example be performed using chemically-synthesized RNA, for which suitable RNA molecules may chemically synthesized using known methods. Alternatively, suitable expression vectors may be used to transcribe such RNA either *in vitro* or *in vivo*. *In vitro* transcription of sense and antisense strands (encoded by sequences present on the same vector or on separate

vectors) may be effected using for example T7 RNA polymerase, in which case the vector may comprise a suitable coding sequence operably-linked to a T7 promoter. The *in vitro*-transcribed RNA may in embodiments be processed (e.g. using *E. coli* RNase III) *in vitro* to a size conducive to RNAi. The sense and antisense transcripts combined to form an RNA duplex which is introduced into a target cell of interest. Other vectors may be used, which express small hairpin RNAs (shRNAs) which can be processed into siRNA-like molecules. Various vector-based methods are known in the art. Various methods for introducing such vectors into cells, either *in vitro* or *in vivo* (e.g. gene therapy) are known in the art.

Accordingly, in an embodiment, expression of a polypeptide including an amino acid sequence substantially identical to a SARS virus sequence may be inhibited by introducing into or generating within a cell an siRNA or siRNA-like molecule corresponding to a nucleic acid molecule encoding the polypeptide or fragment thereof, or to an nucleic acid homologous thereto. In various embodiments such a method may entail the direct administration of the siRNA or siRNA-like molecule into a cell, or use of the vector-based methods described above. In an embodiment, the siRNA or siRNA-like molecule is less than about 30 nucleotides in length. In a further embodiment, the siRNA or siRNA-like molecules are about 21-23 nucleotides in length. In an embodiment, siRNA or siRNA-like molecules comprise and 19-21 bp duplex portion, each strand having a 2 nucleotide 3' overhang. In embodiments, the siRNA or siRNA-like molecule is substantially identical to a nucleic acid encoding the polypeptide or a fragment or variant (or a fragment of a variant) thereof. Such a variant is capable of encoding a protein having the activity of a SARS virus polypeptide. In embodiments, the sense strand of the siRNA or siRNA-like molecule is substantially identical to a SARS virus nucleic acid molecule or a fragment thereof (RNA having U in place of T residues of the DNA sequence).

SARS Virus Protein Expression

In general, SARS virus polypeptides according to the invention, may be produced by transformation of a suitable host cell with all or part of a SARS virus polypeptide-encoding genomic or cDNA molecule or fragment thereof (e.g., the genomic DNA or cDNAs described herein) in a suitable expression vehicle. Those

skilled in the field of molecular biology will understand that any of a wide variety of expression systems may be used to provide the recombinant protein. The precise host cell used is not critical to the invention. The SARS virus polypeptide may be produced in a prokaryotic host (e.g., E. coli or a virus, for example, a coronovirus such as human OC43 or 229E, a bovine coronavirus, or a virus used for gene therapy, such as an adenovirus) or in a eukaryotic host (e.g., Saccharomyces cerevisiae, insect cells, e.g., Sf21cells, or mammalian cells, e.g., COS 1, NIH 3T3, VeroE6, or HeLa cells). Such cells are available from a wide range of sources (e.g., the American Type Culture Collection, Rockland, Md.; also, see, e.g., Ausubel et al., Current Protocols in Molecular Biology, John Wiley & Sons, New York, 1994). The method of transformation or transfection and the choice of expression vehicle will depend on the host system selected. Transformation and transfection methods are described, e.g., in Ausubel et al. (supra); expression vehicles may be chosen from those provided, e.g., in Cloning Vectors: A Laboratory Manual, P. H. Pouwels et al, 1985, Supp. 1987), or from commercially available sources. Suitable animal models, e.g. a ferret animal model, or any other animal model suitable for analysis of SARS virus infection or expression of SARS virus nucleic acid molecules may be used.

In an alternative embodiment, the baculovirus expression system (using, for example, the vector pBacPAK9) available from Clontech (Pal Alto, Calif.) may be used. If desired, this system may be used in conjunction with other protein expression techniques, for example, the myc tag approach described by Evan et al. (Mol. Cell Biol. 5:3610-3616, 1985). In an alternative embodiment, a SARS virus polypeptide may be produced by a stably-transfected mammalian cell line. A number of vectors suitable for stable transfection of mammalian cells are available to the public, e.g., see Pouwels et al (supra); methods for constructing such cell lines are also publicly available, e.g., in Ausubel et al. (supra). In one example, cDNA encoding the SARS virus polypeptide is cloned into an expression vector which includes the dihydrofolate reductase (DHFR) gene. Integration of the plasmid and, therefore, the SARS virus polypeptide-encoding gene into the host cell chromosome is selected for by inclusion of 0.01-300 μM methotrexate in the cell culture medium (as described in Ausubel et al., supra). This dominant selection can be accomplished in most cell types. Recombinant protein expression can be increased by DHFR-mediated amplification of the transfected gene.

Methods for selecting cell lines bearing gene amplifications are described in Ausubel et al. (supra); such methods generally involve extended culture in medium containing gradually increasing levels of methotrexate. DHFR-containing expression vectors commonly used for this purpose include pCVSEII-DHFR and pAdD26SV(A) (described in Ausubel et al., supra). Any of the host cells described above or, preferably, a DHFR-deficient CHO cell line (e.g., CHO DHFR.sup.- cells, ATCC Accession No. CRL 9096) are among the host cells preferred for DHFR selection of a stably-transfected cell line or DHFR-mediated gene amplification.

Once the recombinant SARS virus polypeptide is expressed, it is isolated, e.g., using affinity chromatography. In one example, an anti-SARS virus polypeptide antibody (e.g., produced as described herein) may be attached to a column and used to isolate the SARS virus polypeptide. Lysis and fractionation of SARS virus polypeptde-harboring cells prior to affinity chromatography may be performed by standard methods (see, e.g., Ausubel et al., supra). In another example, SARS virus polypeptides may be purified or substantially purified from a mixture of compounds such as an extract or supernatant obtained from cells (Ausubel et al., supra). Standard purification techniques can be used to progressively eliminate undesirable compounds from the mixture until a single compound or minimal number of effective compounds has been isolated.

Once isolated, the recombinant protein can, if desired, be further purified, e.g., by high performance liquid chromatography (see, e.g., Fisher, Laboratory Techniques In Biochemistry And Molecular Biology, eds., Work and Burdon, Elsevier, 1980).

Polypeptides of the invention, particularly short SARS virus peptide fragments, can also be produced by chemical synthesis (e.g., by the methods described in Solid Phase Peptide Synthesis, 2nd ed., 1984 The Pierce Chemical Co., Rockford, Ill.).

These general techniques of polypeptide expression and purification can also be used to produce and isolate useful SARSvirus protein fragments or analogs (described herein).

In certain alternative embodiments, the SARS polypeptide might have attached any one of a variety of tags. Tags can be amino acid tags or chemical tags and can be added for the purpose of purification (for example a 6-histidine tag for purification over a nickel column). In other preferred embodiments, various labels can be used as means

for detecting binding of a SARS polypeptide to another polypeptide, for example to a cell surface receptor. Alternatively, SARS DNA or RNA may be labeled for detection, for example in a hybridization assay. SARS virus nucleic acids or proteins, or derivatives thereof, may be directly or indirectly labeled, for example, with a radioscope, a fluorescent compound, a bioluminescent compound, a chemiluminescent compound, a metal chelator or an enzyme. Those of ordinary skill in the art will know of other suitable labels or will be able to ascertain such, using routine experimentation. In yet another embodiment of the invention, the polypeptides disclosed herein, or derivatives thereof, are linked to toxins.

Isolation and Identification of Additional SARS virus molecules

Based on the SARS virus sequences described herein, the isolation and identification of additional SARS virus-related sequences such as SARS virus genes and of additional SARS virus strains or isolates is made possible using standard techniques. In addition, the SARS virus sequences provided herein also provide the basis for identification of homologous sequences from other species and genera from both prokaryotes and eukaryotes such as viruses, bacteria, fungi, parasites, yeast, and/or mammals.In some embodiments, the nucleic acid sequences described herein may be used to design probes or primers, including degenerate oligonucleotide probes or primers, based upon the sequence of either DNA strand. The probes or primers may then be used to screen genomic or cDNA libraries for sequences from for example naturally occurring variants or isolates of SARS viruses, using standard amplification or hybridization techniques.

In some embodiments, binding partners may be identified by tagging the polypeptides of the invention (e.g., those substantially identical to SARS virus polypeptides described herein) with an epitope sequence (e.g., FLAG or 2HA), and delivering it into host cells, either by transfection with a suitable vector containing a nucleic acid sequence encoding a polypeptide of the invention, followed by immunoprecipitation and identification of the binding partner. Cells may be infected with strains expressing the FLAG or 2HA fusions, followed by lysis and immunoprecipitation with anti-FLAG or anti-2HA antibodies. Binding partners may be identified by mass spectroscopy . If the polypeptide of the invention is not produced in

sufficient quantities, such a method may not deliver enough tagged protein to identify its partner. As part of a complementary approach, each polypeptide of the invention may be cloned into a mammalian transfection vector fused to, for example, 2HA, GFP and/or FLAG. Following transfection, HeLa cells may be lysed and the tagged polypeptide immunoprecipitated. The binding partner may be identified by SDS PAGE followed by mass spectroscopy.

In some embodiments, polypeptides or antibodies of the invention may be tagged, produced, and used for example on affinity columns and/orin immunological assays to identify and/or confirm identified target compounds. FLAG, HA, and/or His tagged proteins can be used for such affinity columns to pull out host cell factors from cell extracts, and any hits may be validated by standard binding assays, saturation curves, and other methods as described herein or known to those of skill in the art.

In some embodiments, a two hybrid system may be used to study protein-protein interactions. The nucleic acid sequences described herein, or sequences substantially identical thereto, can be cloned into the pBT bait plasmid of the two hybrid system, and a commercially available murine spleen library of 5×10^6 independent clones, may be used as the target library for the baits. Potential hits may be further characterized by recovering the plasmids and retransforming to reduce false positives resulting from clonal bait variants and library target clones which activate the reporter genes independent of the cloned bait. Reproducible hits may be studied further as described herein.

Virulence may be assayed as described herein or as known to those of skill in the art. Once coding sequences have been identified, they may be isolated using standard cloning techniques, and inserted into any suitable vector or replicon for, for example, production of polypeptides. Such vectors and replicons include, without limitation, bacteriophage X (E. coli), pBR322 (E. coli), pACYC177 (E. coli), pKT230 (gram-negative bacteria), pGV1 106 (gram-negative bacteria), pLAFRl (gram-negative bacteria), pME290 (non-E. coli gram-negative bacteria), pHV14 (E. coli and Bacillus subtilis), pBD9 (Bacillus), pIJ61 (Streptomyces), pUC6 (Streptomyces), YIp5 (Saccharomyces), YCpl9 (Saccharomyces) or bovine papilloma virus (mammalian cells). In general, the polypeptides of the invention may be produced in any suitable host cell transformed or transfected with a suitable vector. The method of

transformation or transfection and the choice of expression vehicle will depend on the host system selected. A wide variety of expression systems may be used, and the precise host cell used is not critical to the invention. For example, a polypeptide according to the invention may be produced in a prokaryotic host (e.g., *E. coli*) or in a eukaryotic host (e.g., *Saccharomyces cerevisiae*, insect cells, e.g., Sf21 cells, or mammalian cells, e.g., NIH 3T3, HeLa, or COS cells). Such cells are available from a wide range of sources (e.g., the American Type Culture Collection, Manassus, VA.). Bacterial expression systems for polypeptide production include the E. coli pET expression system (Novagen, Inc., Madison, Wis.), and the pGEX expression system(Pharmacia).

Compounds

In one aspect, compounds according to the invention include SARS virus nucleic acid molecules and polypeptides, such as the sequences disclosed in the Figures and Tables herein, and throughout the specification, and fragments thereof. In alternative embodiments, compounds according to the invention may be nucleic acid molecules that are at least 10 nucleotides in length, and that are derived from the sequences described herein. In alternative embodiments, compounds according to the invention may be peptides that are at least 5 amino acids in length, and that are derived from the sequences described herein.

In alternative embodiments, a compound according to the invention can be a non-peptide molecule as well as a peptide or peptide analogue. A peptide or peptide analogue will generally be as small as feasible while retaining full biological activity. A non-peptide molecule can be any molecule that exhibits biological activity as described herein or known in the art. Biological activity can, for example, be measured in terms of ability to elicit a cytotoxic response, to mediate DNA replication, or any other function of a SARS virus molecule.

Compounds can be prepared by, for example, replacing, deleting, or inserting an amino acid residue of SARS peptide or peptide analogue, as described herein, with other conservative amino acid residues, i.e., residues having similar physical, biological, or chemical properties, and screening for biological function.

It is well known in the art that some modifications and changes can be made in the structure of a polypeptide without substantially altering the biological function of that peptide, to obtain a biologically equivalent polypeptide. Such modifications may be made for the purpose of modifying function, or for facilitating administration or enhancing stability or inhibiting breakdown for, for example, therapeutic uses. For example, an indigestible SARS virus compound according to the invention may be used for oral administration; a modification that is suitable for inhalation may be used for administration to the lung; or addition of a leader sequence may increase protein expression levels.

In one aspect of the invention, SARS virus-derived peptides or epitopes may include peptides that differ from a portion of a native leader, protein or SARS virus sequence by conservative amino acid substitutions. The peptides and epitopes of the present invention also extend to biologically equivalent peptides that differ from a portion of the sequence of novel peptides of the present invention by conservative amino acid substitutions. As used herein, the term "conserved amino acid substitutions" refers to the substitution of one amino acid for another at a given location in the peptide, where the substitution can be made without substantial loss of the relevant function. In making such changes, substitutions of like amino acid residues can be made on the basis of relative similarity of side-chain substituents, for example, their size, charge, hydrophobicity, hydrophilicity, and the like, and such substitutions may be assayed for their effect on the function of the peptide by routine testing.

In some embodiments, conserved amino acid substitutions may be made where an amino acid residue is substituted for another having a similar hydrophilicity value (e.g., within a value of plus or minus 2.0), where the following may be an amino acid having a hydropathic index of about -1.6 such as Tyr (-1.3) or Pro (-1.6)s are assigned to amino acid residues (as detailed in United States Patent No. 4,554,101, incorporated herein by reference): Arg (+3.0); Lys (+3.0); Asp (+3.0); Glu (+3.0); Ser (+0.3); Asn (+0.2); Gln (+0.2); Gly (0); Pro (-0.5); Thr (-0.4); Ala (-0.5); His (-0.5); Cys (-1.0); Met (-1.3); Val (-1.5); Leu (-1.8); Ile (-1.8); Tyr (-2.3); Phe (-2.5); and Trp (-3.4).

In alternative embodiments, conserved amino acid substitutions may be made where an amino acid residue is substituted for another having a similar hydropathic index (e.g., within a value of plus or minus 2.0). In such embodiments, each amino acid

residue may be assigned a hydropathic index on the basis of its hydrophobicity and charge characteristics, as follows: Ile (+4.5); Val (+4.2); Leu (+3.8); Phe (+2.8); Cys (+2.5); Met (+1.9); Ala (+1.8); Gly (-0.4); Thr (-0.7); Ser (-0.8); Trp (-0.9); Tyr (-1.3); Pro (-1.6); His (-3.2); Glu (-3.5); Gln (-3.5); Asp (-3.5); Asn (-3.5); Lys (-3.9); and Arg (-4.5).

In alternative embodiments, conserved amino acid substitutions may be made where an amino acid residue is substituted for another in the same class, where the amino acids are divided into non-polar, acidic, basic and neutral classes, as follows: non-polar: Ala, Val, Leu, Ile, Phe, Trp, Pro, Met; acidic: Asp, Glu; basic: Lys, Arg, His; neutral: Gly, Ser, Thr, Cys, Asn, Gln, Tyr.

Conservative amino acid changes can include the substitution of an L-amino acid by the corresponding D-amino acid, by a conservative D-amino acid, or by a naturally-occurring, non-genetically encoded form of amino acid, as well as a conservative substitution of an L-amino acid. Naturally-occurring non-genetically encoded amino acids include beta-alanine, 3-amino-propionic acid, 2,3-diamino propionic acid, alpha-aminoisobutyric acid, 4-amino-butyric acid, N-methylglycine (sarcosine), hydroxyproline, ornithine, citrulline, t-butylalanine, t-butylglycine, N-methylisoleucine, phenylglycine, cyclohexylalanine, norleucine, norvaline, 2-napthylalanine, pyridylalanine, 3-benzothienyl alanine, 4-chlorophenylalanine, 2-fluorophenylalanine, 3-fluorophenylalanine, 4-fluorophenylalanine, penicillamine, 1,2,3,4-tetrahydro-isoquinoline-3-carboxylix acid, beta-2-thienylalanine, methionine sulfoxide, homoarginine, N-acetyl lysine, 2-amino butyric acid, 2-amino butyric acid, 2,4,-diamino butyric acid, p-aminophenylalanine, N-methylvaline, homocysteine, homoserine, cysteic acid, epsilon-amino hexanoic acid, delta-amino valeric acid, or 2,3-diaminobutyric acid.

In alternative embodiments, conservative amino acid changes include changes based on considerations of hydrophilicity or hydrophobicity, size or volume, or charge. Amino acids can be generally characterized as hydrophobic or hydrophilic, depending primarily on the properties of the amino acid side chain. A hydrophobic amino acid exhibits a hydrophobicity of greater than zero, and a hydrophilic amino acid exhibits a hydrophilicity of less than zero, based on the normalized consensus hydrophobicity scale of Eisenberg *et al.* (*J. Mol. Bio.* 179:125-142, 184). Genetically encoded

hydrophobic amino acids include Gly, Ala, Phe, Val, Leu, Ile, Pro, Met and Trp, and genetically encoded hydrophilic amino acids include Thr, His, Glu, Gln, Asp, Arg, Ser, and Lys. Non-genetically encoded hydrophobic amino acids include t-butylalanine, while non-genetically encoded hydrophilic amino acids include citrulline and homocysteine.

Hydrophobic or hydrophilic amino acids can be further subdivided based on the characteristics of their side chains. For example, an aromatic amino acid is a hydrophobic amino acid with a side chain containing at least one aromatic or heteroaromatic ring, which may contain one or more substituents such as –OH, -SH, -CN, -F, -Cl, -Br, -I, $-NO_2$, -NO, $-NH_2$, -NHR, -NRR, -C(O)R, -C(O)OH, -C(O)OR, $-C(O)NH_2$, -C(O)NHR, -C(O)NRR, etc., where R is independently (C_1-C_6) alkyl, substituted (C_1-C_6) alkyl, (C_1-C_6) alkenyl, substituted (C_1-C_6) alkenyl, (C_1-C_6) alkynyl, substituted (C_1-C_6) alkynyl, (C_5-C_{20}) aryl, substituted (C_5-C_{20}) aryl, (C_6-C_{26}) alkaryl, substituted (C_6-C_{26}) alkaryl, 5-20 membered heteroaryl, substituted 5-20 membered heteroaryl, 6-26 membered alkheteroaryl or substituted 6-26 membered alkheteroaryl. Genetically encoded aromatic amino acids include Phe, Tyr, and Tryp, while non-genetically encoded aromatic amino acids include phenylglycine, 2-napthylalanine, beta-2-thienylalanine, 1,2,3,4-tetrahydro-isoquinoline-3-carboxylic acid, 4-chlorophenylalanine, 2-fluorophenylalanine3-fluorophenylalanine, and 4-fluorophenylalanine.

An apolar amino acid is a hydrophobic amino acid with a side chain that is uncharged at physiological pH and which has bonds in which a pair of electrons shared in common by two atoms is generally held equally by each of the two atoms (i.e., the side chain is not polar). Genetically encoded apolar amino acids include Gly, Leu, Val, Ile, Ala, and Met, while non-genetically encoded apolar amino acids include cyclohexylalanine. Apolar amino acids can be further subdivided to include aliphatic amino acids, which is a hydrophobic amino acid having an aliphatic hydrocarbon side chain. Genetically encoded aliphatic amino acids include Ala, Leu, Val, and Ile, while non-genetically encoded aliphatic amino acids include norleucine.

A polar amino acid is a hydrophilic amino acid with a side chain that is uncharged at physiological pH, but which has one bond in which the pair of electrons shared in common by two atoms is held more closely by one of the atoms. Genetically

encoded polar amino acids include Ser, Thr, Asn, and Gln, while non-genetically encoded polar amino acids include citrulline, N-acetyl lysine, and methionine sulfoxide.

An acidic amino acid is a hydrophilic amino acid with a side chain pKa value of less than 7. Acidic amino acids typically have negatively charged side chains at physiological pH due to loss of a hydrogen ion. Genetically encoded acidic amino acids include Asp and Glu. A basic amino acid is a hydrophilic amino acid with a side chain pKa value of greater than 7. Basic amino acids typically have positively charged side chains at physiological pH due to association with hydronium ion. Genetically encoded basic amino acids include Arg, Lys, and His, while non-genetically encoded basic amino acids include the non-cyclic amino acids ornithine, 2,3,-diaminopropionic acid, 2,4-diaminobutyric acid, and homoarginine.

It will be appreciated by one skilled in the art that the above classifications are not absolute and that an amino acid may be classified in more than one category. In addition, amino acids can be classified based on known behaviour and or characteristic chemical, physical, or biological properties based on specified assays or as compared with previously identified amino acids. Amino acids can also include bifunctional moieties having amino acid-like side chains.

Conservative changes can also include the substitution of a chemically derivatised moiety for a non-derivatised residue, by for example, reaction of a functional side group of an amino acid. Thus, these substitutions can include compounds whose free amino groups have been derivatised to amine hydrochlorides, p-toluene sulfonyl groups, carbobenzoxy groups, t-butyloxycarbonyl groups, chloroacetyl groups or formyl groups. Similarly, free carboxyl groups can be derivatized to form salts, methyl and ethyl esters or other types of esters or hydrazides, and side chains can be derivatized to form O-acyl or O-alkyl derivatives for free hydroxyl groups or N-im-benzylhistidine for the imidazole nitrogen of histidine. Peptide analogues also include amino acids that have been chemically altered, for example, by methylation, by amidation of the C-terminal amino acid by an alkylamine such as ethylamine, ethanolamine, or ethylene diamine, or acylation or methylation of an amino acid side chain (such as acylation of the epsilon amino group of lysine). Peptide analogues can also include replacement of the amide linkage in the peptide with a substituted amide

(for example, groups of the formula –C(O)-NR, where R is (C_1-C_6) alkyl, (C_1-C_6) alkenyl, (C_1-C_6) alkynyl, substituted (C_1-C_6) alkyl, substituted (C_1-C_6) alkenyl, or substituted (C_1-C_6) alkynyl) or isostere of an amide linkage (for example, -CH_2NH-, -CH_2S, -CH_2CH_2-, -CH=CH- (cis and trans), -$C(O)CH_2$-, -$CH(OH)CH_2$-, or –CH_2SO-).

The compound can be covalently linked, for example, by polymerisation or conjugation, to form homopolymers or heteropolymers. Spacers and linkers, typically composed of small neutral molecules, such as amino acids that are uncharged under physiological conditions, can be used. Linkages can be achieved in a number of ways. For example, cysteine residues can be added at the peptide termini, and multiple peptides can be covalently bonded by controlled oxidation. Alternatively, heterobifunctional agents, such as disulfide/amide forming agents or thioether/amide forming agents can be used. The compound can also be constrained, for example, by having cyclic portions.

In some embodiments, three dimensional molecular modeling techniques may be used to identify or generate compounds that may be useful as therapeutics or diagnostics. Standard molecular modeling tools may be used, for example, those described in L-H Hung and R. Samudrala, PROTINFO: secondary and tertiary protein structure prediction, Nucleic Acids Research, 2003, Vol. 31, No. 13 3296-3299; A. Yamaguchi, et al. , Enlarged FAMSBASE: protein 3D structure models of genome sequences for 41 species, Nucleic Acids Research, 2003, Vol. 31, No. 1 463-468; J. Chen, et al., MMDB: Entrez's 3D-structure database,Nucleic Acids Research, 2003, Vol. 31, No. 1 474-477; R. A. Chiang, et al., The Structure Superposition Database, Nucleic Acids Research, 2003, Vol. 31, No. 1 505-510.

Peptides or peptide analogues can be synthesized by standard chemical techniques, for example, by automated synthesis using solution or solid phase synthesis methodology. Automated peptide synthesizers are commercially available and use techniques well known in the art. Peptides and peptide analogues can also be prepared using recombinant DNA technology using standard methods such as those described in, for example, Sambrook, *et al.* (Molecular Cloning: A Laboratory Manual. 2.sup.nd, ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y., 1989) or Ausubel *et al.* (Current Protocols in Molecular Biology, John Wiley & Sons, 1994).

Compounds, such as peptides (or analogues thereof) can be identified by routine experimentation by, for example, modifying residues within SARS peptides; introducing single or multiple amino acid substitutions, deletions, or insertions, and identifying those compounds that retain biological activity, *e.g.*, those compounds that have cytotoxic ability.

In general, candidate compounds for prevention or treatment of SARS virus-mediated disorders are identified from large libraries of both natural product or synthetic (or semi-synthetic) extracts or chemical libraries according to methods known in the art. Candidate or test compounds may include, without limitation, peptides, polypeptides, synthesised organic molecules, naturally occurring organic molecules, and nucleic acid molecules. In some embodiments, such compounds screen for the ability to inhibit SARS virus replication or pathogenicity, while maintaining the infected cell's ability to grow or survive.

Those skilled in the field of drug discovery and development will understand that the precise source of test extracts or compounds is not critical to the method(s) of the invention. Accordingly, virtually any number of chemical extracts or compounds can be screened using the exemplary methods described herein or using standard methods. Examples of such extracts or compounds include, but are not limited to, plant-, fungal-, prokaryotic- or animal-based extracts, fermentation broths, and synthetic compounds, as well as modification of existing compounds. Numerous methods are also available for generating random or directed synthesis (e.g., semi-synthesis or total synthesis) of any number of chemical compounds, including, but not limited to, saccharide-, lipid-, peptide-, and nucleic acid-based compounds. Synthetic compound libraries are commercially available. Alternatively, libraries of natural compounds in the form of bacterial, fungal, plant, and animal extracts are commercially available from a number of sources, including Biotics (Sussex, UK), Xenova (Slough, UK), Harbor Branch Oceanographic Institute (Ft. Pierce, Fla.), and PharmaMar, U.S.A. (Cambridge, Mass.). In addition, natural and synthetically produced libraries of, for example, SARS virus polypeptides containing leader sequences, are produced, if desired, according to methods known in the art, e.g., by standard extraction and fractionation methods. Furthermore, if desired, any library or compound is readily modified using standard chemical, physical, or biochemical methods.

When a crude extract is found to modulate cytotoxicity or viral infection, further fractionation of the positive lead extract is necessary to isolate chemical constituents responsible for the observed effect. Thus, the goal of the extraction, fractionation, and purification process is the careful characterization and identification of a chemical entity within the crude extract having, for example, anti-cytotoxicity or anti-viral properties. The same assays described herein for the detection of activities in mixtures of compounds can be used to purify the active component and to test derivatives thereof. Methods of fractionation and purification of such heterogenous extracts are known in the art. If desired, compounds shown to be useful agents for treatment are chemically modified according to methods known in the art. Compounds identified as being of therapeutic, prophylactic, diagnostic, or other value in for example cell culture systems, such as a Vero E6 culture system, may be subsequently analyzed using a ferret animal model, or any other animal model suitable for analysis of SARS.

Antibodies

The compounds of the invention can be used to prepare antibodies to SARS virus peptides, protein, polyproteins, or analogs thereof, or to SARS virus nucleic acid molecules or analogs thereof using standard techniques of preparation as, for example, described in Harlow and Lane (Antibodies; A Laboratory Manual, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., 1988), or known to those skilled in the art. Antibodies may include polyclonal antibodies, monoclonal antibodies, hybrid antibodies (e.g., divalent antibodies having different pairs of heavy and light chains), chimeric antibodies (e.g., antibodies having constant and variable domains from different species and/or class), modified antibodies (e.g, antibodies in which the naturally occurring sequence has been altered by for example recombinant techniques), Fab antibodies, anti-idiotype antibodies, etc. Antibodies can be tailored to minimise adverse host immune response by, for example, using chimeric antibodies containing an antigen binding domain from one species and the Fc portion from another species, or by using antibodies made from hybridomas of the appropriate species. For example, "humanized" antibodies may be used for administration to humans.

To generate SARS virus polypeptide-specific antibodies, a SARS virus polypeptide coding sequence may be expressed, for example, as a C-terminal fusion with glutathione S-transferase (GST) (Smith et al., Gene 67:31-40, 1988). The fusion polypeptide may then be purified on glutathione-Sepharose beads, eluted with glutathione cleaved with thrombin (at the engineered cleavage site), and purified to the degree necessary for immunization of rabbits. Primary immunizations are carried out with Freud's complete adjuvant and subsequent immunizations with Freud's incomplete adjuvant. Antibody titres are monitored by Western blot and immunoprecipitation analyzes using the thrombin-cleaved SARS virus polypeptide fragment of the GST-SARS virus fusion polypeptide. Immune sera are affinity purified using CNBr-Sepharose-coupled SARS virus polypeptide. Antiserum specificity is determined using a panel of unrelated GST polypeptides.

As an alternate or adjunct immunogen to GST fusion polypeptides, peptides corresponding to relatively unique hydrophilic SARS virus polypeptides may be generated and coupled to keyhole limpet hemocyanin (KLH) through an introduced C-terminal lysine. Antiserum to each of these peptides is similarly affinity purified on peptides conjugated to BSA, and specificity tested in ELISA and Western blots using peptide conjugates, and by Western blot and immunoprecipitation using SARS virus polypeptide expressed as a GST fusion polypeptide.

Alternatively, monoclonal antibodies may be prepared using the SARS virus polypeptides described above and standard hybridoma technology (see, e.g., Kohler et al., Nature, 256:495, 1975; Kohler et al., Eur. J Immunol. 6:511, 1976; Kohler et al., Eur. J. Immunol. 6:292, 1976; Hammerling et al., In Monoclonal Antibodies and T Cell Hybridomas, Elsevier, NY, 1981; Ausubel et al., supra). Once produced, monoclonal antibodies are also tested for specific SARS virus polypeptide recognition by Western blot or immunoprecipitation analysis (by the methods described in Ausubel et al., supra). Antibodies which specifically recognize SARS virus polypeptides are considered to be useful in the invention; such antibodies may be used, e.g., in an immunoassay to monitor the level of SARS virus polypeptides produced by a mammal (for example, to determine the amount or location of a SARS virus polypeptide).

In an alternative embodiment, antibodies of the invention are not only produced using the whole SARS virus polypeptide, but using fragments of the SARS virus

polypeptide which are unique or which lie outside highly conserved regions and appear likely to be antigenic, by criteria such as high frequency of charged residues may also be used. In one specific example, such fragments are generated by standard techniques of PCR and cloned into the pGEX expression vector (Ausubel et al., supra). Fusion polypeptides are expressed in E. coli and purified using a glutathione agarose affinity matrix as described in Ausubel et al. (supra). To attempt to minimize the potential problems of low affinity or specificity of antisera, two or three such fusions are generated for each polypeptide, and each fusion is injected into at least two rabbits. Antisera are raised by injections in a series, preferably including at least three booster injections. SARS virus antibodies may also be prepared against SARS virus nucleic acid molecules.

Antibodies may be used as diagnostics, therapeutics, or prophylactics for SARS virus-related disorders. Antibodies may also be used to isolate SARS virus and compounds by for example affinity chromatography, or to identify SARS virus compounds isolated or generated by other techniques.

Arrays and Libraries

In some aspects, biological assays, such as diagnostic or other assays, using high density nucleic acid, polypeptide, or antibody arrays, for example high density miniaturized arrays or "microarrays," of SARS virus nucleic acid molecules or polypeptides, or antibodies capable of specifically binding such nucleic acid molecules or polypeptides, may be performed. Macroarrays, performed for example by manual spotting techniques, may also be used. Arrays generally require a solid support (for example, nylon, glass, ceramic, plastic, silicon, nitrocellulose or PVDF membranes, microwells, microbeads, e.g., magnetic microbeads, etc.) to which the nucleic acid molecules or polypeptides or antibodies are attached in a specified two-dimensional arrangement, such that the pattern of hybridization is easily determinable. Suspension arrays (particles in suspension) that are coded to facilitate identification may also be used. SARS virus nucleic acid molecules or polypeptide probes or targets may be compounds as described herein.

In some embodiments, high density nucleic acid arrays may for example be used to monitor the presence or level of expression of a large number of SARS virus

nucleic acid molecules or genes or for detecting or identifying SARS virus nucleic acid sequence variations, mutations or polymorphisms. For the purpose of such arrays, "nucleic acids" may include any polymer or oligomer of nucleosides or nucleotides (polynucleotides or oligonucleotides), which include pyrimidine and purine bases, preferably cytosine, thymine, and uracil, and adenine and guanine, respectively, or may include peptide nucleic acids (PNA). In an alternative aspect, the invention provides nucleic acid microarrays including a number of distinct nucleic acid sequence arrays of the invention, thus providing specific "sets" of sequences. The number of distinct sequences may for example be any integer between 2 and 1 x 10^5, such as at least 10^2, 10^3, 10^4, or 10^5.

The invention also provides gene knockout and expression libraries. Thus, nucleic acid molecules encoding SARS virus polypeptides or proteins (e.g., PCR products of ORF's or total mRNA) may for example be attached to a solid support, hybridized with single stranded detectably-labeled cDNAs (corresponding to an "antisense" orientation), and quantified using an appropriate method such that a signal is detected at each location at which hybridization has taken place. The intensity of the signal would then reflect the level of gene expression. Comparison of results from viruses, for example, of different strains or from different samples or subjects, would elucidate differing levels of expression of specified genes. Using similar techniques, homologous nucleic acids may be identified from different viruses if SARS virus nucleic acids are used in the microarray, and probed with nucleic acid molecules from different viruses or subjects. In some embodiments, this approach may involve constructing his-tagged ORF expression libraries of viral genomes in a bacterial host, similar to an expression library in yeast (Martzen M. R. et al., 1999. Science, 286:1153). ORF-encoded protein activities may for example be detected in purified his-tagged protein pools in cases where activities cannot be detected in extracts or cells. In one aspect of the invention, arrayed libraries may be constructed of viral strains each of which bears a plasmid expressing a different SARS virus ORF under control of an inducible promoter. ORFs are amplified using PCR and cloned into a vector that enables their expression as N-terminal his-tagged polypeptides. These amplicons are also used to construct hybridization microarrays and enable targeted gene disruption, reducing expenses. A suitable expression host is selected, and genes encoding

particular biochemical activities are identified by screening arrayed pools of his-tagged proteins as described previously (Martzen M. R., McCraith S. M., Spinelli S.L., Torres F. M., Fields S., Grayhack E.J., and Phizicky E. M., 1999. Science, 286:1153).

In some embodiments, protein arrays (including antibody or antigen arrays) may be used for the analysis and identification of SARS virus polypeptides or host responses to such polypeptides. Thus, protein arrays may be used to detect SARS virus polypeptides in a patient; distinguish a SARS virus polypeptide from a host polypeptide; detect interactions between SARS virus polypeptides and for example host proteins; determine the efficacy of potential therapeutics, such as small molecules or ligands that may bind SARS virus polypeptides; determine protein-antibody interactions; and/or detect the interaction of enzyme-substrate interactions. Protein arrays may also be used to detect SARS virus antigens and antibodies in samples; to profile expression of SARS virus polypeptides; to identify suitable antibodies or map epitopes; or for a variety of protein function analyses.

A variety of methods are known for making and using microarrays, as for example disclosed in Cheung V. G., *et al.*, 1999. Nature Genetics Supplement, 21:15-19; Lipshutz R. J., *et al.*,1999. Nature Genetics Supplement, 21:20-24; Bowtell D. D. L., 1999. Nature Genetics Supplement, 21:25-32; Singh-Gasson S., *et al.*, 1999. Nature Biotechnol., 17:974-978; and Schweitzer B., *et al.*, 2002. Nature Biotechnol., 20:359-365. Thus, for example, microarrays may be designed by synthesizing oligonucleotides with sequence variations based on a reference sequences, such as any SARS virus sequences described herein. Methods for storing, querying and analyzing microarray data have for example been disclosed in, for example, United States Patent No. 6,484,183; United States Patent No. 6,188,783; and Holloway A. J., *et al.*, 2002. Nature Genetics Supplement, 32:481-489. Protein arrays may be constructed, detected, and analysed using methods known in the art for example mass spectrometric techniques, immunoassays such as ELISA and western (dot) blotting combined with for example fluorescence detection techniques, and adapted for high throughput analysis, as described in for example MacBeath, G. and Schreiber, S.L. Science 2000, 289, 1760-1763; Levit-Binnun N, et al. (2003) Quantitative detection of protein arrays. Anal Chem 75:1436-41; Kukar T, et al. (2002) Protein microarrays to detect protein-protein interactions using red and green fluorescent proteins. Anal Biochem 306:50-4;

Borrebaeck CA, et al. (2001) Protein chips based on -recombinant antibody fragments: a highly sensitive approach as detected by mass spectrometry. Biotechniques 30:1126-1132; Huang RP (2001) Detection of multiple proteins in an antibody-based protein microarray system. J Immunol Methods 255:1-13; Emili AQ and Cagney G (2000) Large-scale functional analysis using peptide or protein arrays. Nature Biotechnol 18:393-397; Zhu H, et al. (2000) Analysis of yeast protein kinases using protein chips. Nature Genet 26:283-9; Lueking A, et al. (1999) Protein Microarrays for Gene Expression and Antibody Screening. Anal. Biochem. 270:103-111; or Templin MF, et al. (2002) Protein microarray technology. Drug Discov Today 7:815-822. Tools for microarray techniques are available commercially from for example Affymetrix, Santa Clara, CA; Nanogen, San Diego, CA; or Sequenom, San Diego, CA.

Computer Readable Records

Nucleic acid and polypeptide sequences, as described herein, or a fragment thereof, may be provided in a variety of media to facilitate access to these sequences and enable the use thereof. According, SARS virus nucleic acid and polypeptide sequences of the invention may be recorded or stored on computer readable media, using any technique and format that is appropriate for the particular medium.

In alternative embodiments, the invention provides computer readable media encoded with a number of distinct nucleic acid or amino acid data sequences of the invention. The number of distinct sequences may for example be any integer between 2 and 1 x 10^5, such as at least 10^2, 10^3, 10^4, or 10^5. In one embodiment, the invention features a computer medium having a plurality of digitally encoded data records. Each data record may include a value representing a nucleic acid or amino acid sequence of the invention. In some embodiments, the data record may further include values representing the level of expression, level or activity of a nucleic acid or amino acid sequence of the invention. The data record can be structured as a table, for example, a table that is part of a database such as a relational database (for example, a SQL database of the Oracle or Sybase database environments). The invention also includes a method of communicating information about a sample, for example by transmitting information, for example transmitting a computer readable record as described herein, for example over a computer network. The polypeptide and nucleic acid sequences of

the invention, and sequence information pertaining thereto, may be routinely accessed by one of ordinary skill in the art for a variety of purposes, including for the purposes of comparing substantially identical sequences, etc. Such access may be facilitated using publicly available software as described herein. By "computer readable media" is meant any medium that can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media.

Pharmaceutical and Veterinary Compositions, Dosages, And Administration

Compounds of the invention can be provided alone or in combination with other compounds (for example, small molecules, peptides, or peptide analogues), in the presence of a liposome, an adjuvant, or any pharmaceutically acceptable carrier, in a form suitable for administration to humans or to animals.

Conventional pharmaceutical practice may be employed to provide suitable formulations or compositions to administer the compounds to patients suffering from or presymptomatic for SARS. Any appropriate route of administration may be employed, for example, parenteral, intravenous, subcutaneous, intramuscular, intracranial, intraorbital, ophthalmic, intraventricular, intracapsular, intraspinal, intracisternal, intraperitoneal, intranasal, aerosol, or oral administration. In some embodiments, compounds are delivered directly to the lung, by for example, formulations suitable for inhalation. In some embodiments, gene therapy techniques may be used for administration of SARS virus nucleic acid molecules, for example, as DNA vaccines.Formulations may be in the form of liquid solutions or suspensions; for oral administration, formulations may be in the form of tablets or capsules; and for intranasal formulations, in the form of powders, nasal drops, or aerosols.

Methods well known in the art for making formulations are found in, for example, "Remington's Pharmaceutical Sciences" (18th edition), ed. A. Gennaro, 1990, Mack Publishing Company, Easton, Pa. Formulations for parenteral administration may, for example, contain excipients, sterile water, or saline, polyalkylene glycols such as polyethylene glycol, oils of vegetable origin, or hydrogenated napthalenes.

Biocompatible, biodegradable lactide polymer, lactide/glycolide copolymer, or polyoxyethylene-polyoxypropylene copolymers may be used to control the release of the compounds. Other potentially useful parenteral delivery systems for modulatory compounds include ethylene-vinyl acetate copolymer particles, osmotic pumps, implantable infusion systems, and liposomes. Formulations for inhalation may contain excipients, for example, lactose, or may be aqueous solutions containing, for example, polyoxyethylene-9-lauryl ether, glycocholate and deoxycholate, or may be oily solutions for administration in the form of nasal drops, or as a gel.

If desired, treatment with a compound according to the invention may be combined with more traditional therapies for the disease.

For therapeutic or prophylactic compositions, the compounds are administered to an individual in an amount sufficient to stop or slow the replication of the SARS virus, or to confer protective immunity against future SARS virus infection. Amounts considered sufficient will vary according to the specific compound used, the mode of administration, the stage and severity of the disease, the age, sex, and health of the individual being treated, and concurrent treatments. As a general rule, however, dosages can range from about 1μg to about 100 mg per kg body weight of a patient for an initial dosage, with subsequent adjustments depending on the patient's response, which can be measured, for example by determining the presence of SARS nucleic acid molecules, polypeptides, or virions in the patient's peripheral blood.

In the case of vaccine formulations, an immunogenically effective amount of a compound of the invention can be provided, alone or in combination with other compounds, with an adjuvant, for example, Freund's incomplete adjuvant or aluminum hydroxide. The compound may also be linked with a carrier molecule, such as bovine serum albumin or keyhole limpet hemocyanin to enhance immunogenicity.
In general, compounds of the invention should be used without causing substantial toxicity. Toxicity of the compounds of the invention can be determined using standard techniques, for example, by testing in cell cultures or experimental animals and determining the therapeutic index, i.e., the ratio between the LD50 (the dose lethal to 50% of the population) and the LD100 (the dose lethal to 100% of the population). In some circumstances however, such as in severe disease conditions, it may be necessary to administer substantial excesses of the compositions.

Virus Isolation

Virus isolation was performed on a bronchoaveolar lavage specimen of a fatal SARS case belonging to the original case cluster from Toronto, Canada. All work with the infectious agent was performed in a biosafety level 3 (BSL3) laboratory using a N100 mask for personal protection. Samples were removed from BSL3 after addition of the RNA extraction buffer. The virus isolate, named the "Tor2 isolate" was grown in African Green Monkey Kidney (Vero E6) cells, the viral particles were purified, and the genetic material (RNA) was extracted from the Tor2 isolate (Poutanen, S. M. et al., N Engl J Med, Apr 10, 2003). More specifically, one hundred microlitre specimens were used to inoculate Vero E6 cells (ATCC CRL 1586) on Dulbecco's Modified Eagle Medium supplemented with penicillin/ streptomycin, glutamine and 2% fetal calf serum. The culture was incubated at 37°C. Cytopathogenic effect was observed 5 days post inoculation. The virus was passaged into newly seeded Vero E6 cells which showed a cytopathogenic effect as early as 2 days post infection (multiplicity of infection 10^{-2}). A virus stock was prepared from passage 2 of these cells and preserved in liquid nitrogen. The titer of the virus stock was determined to be $1x10^7$ plaque forming units (p.f.u.) by plaque assay and 5×10^6 by tissue culture infectious dose (TCID) 50.

For virus propagation, 10 x T-162 flasks of Vero E6 cells were infected with a multiplicity of infection of 10^{-2}. When infected cells showed a cytopathognic effect of '4+' (48 hours post infection), the cultures were then frozen and thawed to lyse the cells, and the supernatants were clarified from cell debris by centrifugation at 10,000 rpm in a Beckman high-speed centrifuge. The supernatants were treated with DNAse and RNAse for 3 hours at 37°C to remove any cellular genomic nucleic acids and subsequently extracted with an equal volume of 1,1,2-trichloro-trifluoroethane. The top fraction was ultra-centrifuged through a 5% / 40% glycerol step gradient at 151,000 x g for 1 hour at 4°C. The virus pellet was resuspended in PBS. RNA was isolated using a commercial kit from QIAGEN and stored at -80°C for further use.

cDNA Library Construction

The RNA and subsequent products were handled under biosafety level 2 (BSL2) conditions. The RNA sample was converted to a cDNA library, using a combined random-priming and oligo-dT priming strategy, and resultant subgenomic clones were processed under level 1 biosafety conditions. More specifically, purified viral RNA (55 ng) was used in the construction of a random primed and oligo-dT primed cDNA library, using the SuperScript Choice System for cDNA synthesis (Invitrogen). Linkers 5' –AATTCGCGGCCGCGTCGAC-3', SEQ ID NO: 195, and 5'-pGTCGACGCGGCCGCG-3', SEQ ID NO: 196, were ligated following cDNA synthesis. The cDNA synthesis products were visualized on agarose gels, revealing the anticipated low-yield smear. To produce sufficient cDNA for cloning, the cDNA product was size fractionated on a low-melting point preparative agarose gel, followed by PCR amplification using a single PCR primer 5'AATTCGCGGCCGCGTCGAC-3', SEQ ID NO: 197, specific to the linkers. This yielded sufficient material for cloning.

Size-selected cDNA products were cloned and single sequence reads were generated from each end of the insert from randomly picked clones. A list of the SARS virus clones is provided in the accompanying sequence listing, which is incorporated by reference herein (SEQ ID NOs: 92-159, 208 and 209).

More specifically, size-selected cDNAs were ligated into the pCR4-TOPO TA cloning vector (Invitrogen, CA), or after digestion with the restriction nuclease Not I into the pBR194c vector (The Institute for Genomic Research, Rockville, MD, USA). Ligated clones were then transformed by electroporation into DH10B T1 cells (Invitrogen), plated on 22 cm agar plates with the appropriate antibiotic and grown for 16 hours at 37°C. Colonies were picked into 384-well Axygen culture blocks containing 2 X YT media and grown in a shaking incubator for 18 hours at 37°C. Cells were lysed and DNA purified using standard laboratory procedures. Sequencing primers for the 194c clones were 5'-GGCCTCTTCGCTATTACGC-3' (forward primer) and 5' TGCAGGTCGACTCTAGAGGAT-3' (reverse primer).

DNA Sequencing And Assembly Of Reads

Sequences were assembled and the assembly edited to produce the genomic sequence of the SARS virus. More specifically, DNA sequencing of both ends of the plasmid templates was achieved using Applied Biosystems BigDye terminator reagent

(version 3), with electrophoresis and data collection on AB 3700 and 3730 XL instruments DNA sequence reads were screened for non-viral contaminating sequences, trimmed for quality using PHRED (Ewing, B, and P. Green, *Genome Res* 8, 186-94, Mar, 1998) and assembled using PHRAP (Gordon, D. et al. *Genome Res* 8, 195-202, Mar, 1998). Simultaneously, sequences were used in BLAST searches of viral nucleotide and non-redundant protein datasets (NCBI, National Library of Medicine) to search for similarities. Sequence assemblies were visualized using CONSED (Gordon, D. et al. *Genome Res* 8, 195-202, Mar, 1998). Sequence mis-assemblies and contig joins were identified using Miropeats (Parsons, J. D., *Comput Appl Biosci* 11, 615-9 (Dec, 1995). As sequence data accrued, the additional sequences were assembled until it became apparent that the additional depth of sampling was increasing depth of coverage but not extending the length of the contig. At this point, 3,080 sequencing reads were generated, 2,634 of which were assembled into a single large contig.

The sequence information was imported into an ACEDB database (Durbin, J. Thierry-Mieg. 1991-. A *C. elegans* Database. Documentation, code and data available from anonymous FTP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov) and subjected to biological analysis including the identification of open reading frames, detection of similar sequences by BLAST and searching for apparent frameshifts. When frameshifts were identified by this analysis, the sequence assembly was consulted for evidence of sequencing errors and if found, they were corrected. The sequences were also searched for any that could extend the 5' end of the sequence and these were incorporated when found. High quality sequence discrepancies between different sequence reads were identified and resolved. Sequence reads classified as deleted or chimeric were identified through manual inspection and removed from the assembly. The resulting sequence has an average PHRED consensus quality score of 89.96. The lowest quality bases in the assembly are in the immediate vicinity of the 5' and 3' ends of the viral genome, with the lowest quality base having a PHRED score of 35. Most (29,694 of the 29,736 (99.86%)) of the bases have a consensus score of 90. Almost all regions of the genome are represented by reads derived from both strands of the plasmid sequencing templates, the exceptions being 50 bases at the 5' end represented by a single sequencing read, and 5 bases at the 3' end

represented by a single read. The average base in the assembly is represented by 30 reads in the forward direction and 30 reads in the reverse direction, as determined by PHRED. RT-PCR products predicted from the sequence and spanning the entire genome yield PCR products of the anticipated size on agarose gels. To confirm the 5' end of the viral genome RACE was performed using the RLM-RACE kit from Ambion, and primers 5'-CAGGAAACAGCTATGACACCAAGAACAAGGCTCTCCA-3' (SEQ ID NO: 90) and 5'-CAGGAAACAGCTATGACGATAGGGCCTCTTCCACAGA-3' (SEQ ID NO: 91). Fourteen clones were recovered and sequenced. Analysis of these sequences confirmed the 5' end of the coronavirus genome. The SARS genomic sequences have been deposited into Genbank (Accession Nos. AY274119.1, AY274119.2, and AY274119.3).

While the invention has been described in connection with specific embodiments thereof, it will be understood that it is capable of further modifications and this application is intended to cover any variations, uses, or adaptations of the invention following, in general, the principles of the invention and including such departures from the present disclosure that come within known or customary practice within the art to which the invention pertains, and may be applied to the essential features set forth herein and in the scope of the appended claims.

All patents, patent applications, and publications referred to herein are hereby incorporated by reference in their entirety to the same extent as if each individual patent, patent application, or publication was specifically and individually indicated to be incorporated by reference in its entirety.